

Claude Elwood Shannon Miscellaneous Writings

Edited by

N. J. A. Sloane
Aaron D. Wyner

Back in 1993, the late Aaron Wyner and I edited Claude Elwood Shannon's papers, and most of them appeared in a volume (Claude Elwood Shannon's Collected Papers) which was published by the IEEE Press.

However, there were a number of items written by Shannon of lesser interest which we did not include (some declassified wartime memoranda, obscure AT&T Bell Labs memos, some mimeographed MIT lecture notes, etc.).

These we put into a binder, held together by an Acco metal strip. We made half a dozen copies, and gave copies to the Library of Congress, the British Library, the Bell Laboratories Library, the MIT Library, to Claude Shannon himself, and to one or two other places.

Over the years many people have asked me if it was possible to get access to this collection.

I had now had this volume scanned and converted to pdf files. The total size of the files is about 450 megabytes.

Neil J. A. Sloane, October 13, 2013

CONTENTS

File 1: Front matter

This volume contains the following items. Bracketed numbers refer to the bibliography.

- File 5: [5] “The Use of the Lakatos-Hickman Relay in a Subscriber Sender,” Memorandum MM 40-130-179, August 3, 1940, Bell Laboratories, 7 pp. + 8 figs.
- File 7: [7] “A Study of the Deflection Mechanism and Some Results on Rate Finders,” Report to National Defense Research Committee, Div. 7-311-M1, circa April, 1941, 37 pp. + 15 figs.
- File 9: [9] “A Height Data Smoothing Mechanism,” Report to National Defense Research Committee, Div. 7-313.2-M1, Princeton Univ., May 26, 1941, 9 pp. + 9 figs.
- File 11: [11] “Some Experimental Results on the Deflection Mechanism,” Report to National Defense Research Committee, Div. 7-311-M1, June 26, 1941, 11 pp.
- File 12: [12] “Criteria for Consistency and Uniqueness in Relay Circuits,” Typescript, Sept. 8, 1941, 5 pp. + 3 figs.
- File 16: [16] (With W. Feller) “On the Integration of the Ballistic Equations on the Aberdeen Analyzer,” Applied Mathematics Panel Report No. 28.1, National Defense Research Committee, July 15, 1943, 9 pp.
- File 16: [19] “Two New Circuits for Alternate Pulse Counting,” Typescript, May 29, 1944, Bell Laboratories, 2 pp. + 3 Figs.

(Note that many of these files contain more than one document.)

File 16:
File 21:
File 21:
File 24:
File 26:
File 27:

File 30:
File 31:
File 31:
File 31:
File 31:
File 36:
File 36:
File 46:
File 46:
File 46:

- [20] "Counting Up or Down With Pulse Counters," Typescript, May 31, 1944, Bell Laboratories, 1 p. + 1 fig.
- [21] (With B. M. Oliver) "Circuits for a P.C.M. Transmitter and Receiver," Memorandum MM 44-110-37, June 1, 1944, Bell Laboratories, 4 pp., 11 figs.
- [23] "Pulse Shape to Minimize Bandwidth With Nonoverlapping Pulses," Typescript, August 4, 1944, Bell Laboratories, 4 pp.
- [24] "A Mathematical Theory of Cryptography," Memorandum MM 45-110-02, Sept. 1, 1945, Bell Laboratories, 114 pp. + 25 figs.
- [26] "Mixed Statistical Determinate Systems," Typescript, Sept. 19, 1945, Bell Laboratories, 17 pp.
- [27] (With R. B. Blackman and H. W. Bode) "Data Smoothing and Prediction in Fire-Control Systems," Summary Technical Report, Div. 7, National Defense Research Committee, Vol. 1, *Gunfire Control*, Washington, DC, 1946, pp. 71-159 and 166-167. AD 200795. Also in National Military Establishment Research and Development Board, Report #13 MGC 12/1, August 15, 1948. Superseded by [51] and by R. B. Blackman, *Linear Data-Smoothing and Prediction in Theory and Practice*, Addison-Wesley, Reading, Mass., 1965.
- [30] (With C. L. Dolph) "The Transient Behavior of a Large Number of Four-Terminal Unilateral Linear Networks Connected in Tandem," Memorandum MM 46-110-49, April 10, 1946, Bell Laboratories, 34 pp. + 16 figs.
- [31] "Electronic Methods in Telephone Switching," Typescript, October 17, 1946, Bell Laboratories, 5 pp. + 1 fig.
- [32] "Some Generalizations of the Sampling Theorem," Typescript, March 4, 1948, 5 pp. + 1 fig.
- [34] "The Normal Ergodic Ensembles of Functions," Typescript, March 15, 1948, 5 pp.
- [35] "Systems Which Approach the Ideal as $P/N \rightarrow \infty$," Typescript, March 15, 1948, 2 pp.
- [36] "Theorems on Statistical Sequences," Typescript, March 15, 1948, 8 pp.
- [45] "Significance and Application [of Communication Research]," *Symposium on Communication Research, 11-13 October, 1948*, Research and Development Board, Department of Defense, Washington, DC, pp. 14-23, 1948.
- [46] "Note on Certain Transcendental Numbers," Typescript, October 27, 1948, Bell Laboratories, 1 p.
- [47] "A Case of Efficient Coding for a Very Noisy Channel," Typescript, Nov. 18, 1948, Bell Laboratories, 2 pp.
- [48] "Note on Reversing a Discrete Markhoff Process," Typescript, Dec. 6 1948, Bell Laboratories, 2 pp. + 2 Figs.

File 46:

[49] "Information Theory," Typescript of abstract of talk for American Statistical Society, 1949, 5 pp.

File 46:

[58] "Proof of an Integration Formula," Typescript, circa 1950, Bell Laboratories, 2 pp.

File 59:

[59] "A Digital Method of Transmitting Information," Typescript, no date, circa 1950, Bell Laboratories, 3 pp.

File 59:

[72] "Creative Thinking," Typescript, March 20, 1952, Bell Laboratories, 10 pp.

File 59:

[74] (With E. F. Moore) "The Relay Circuit Analyzer," Memorandum MM 53-1400-9, March 31, 1953, Bell Laboratories, 14 pp. + 4 figs.

File 59:

[77] "Throbac - Circuit Operation," Typescript, April 9, 1953, Bell Laboratories, 7 pp.

File 78:

[78] "Tower of Hanoi," Typescript, April 20, 1953, Bell Laboratories, 4 pp.

File 78:

[81] "Mathmanship or How to Give an Explicit Solution Without Actually Solving the Problem," Typescript, June 3, 1953, Bell Laboratories, 2 pp.

File 78:

[84] (With E. F. Moore) "The Relay Circuit Synthesizer," Memorandum MM 53-140-52, November 30, 1953, Bell Laboratories, 22 pp. + 5 figs.

File 78:

[87] "Bounds on the Derivatives and Rise Time of a Band and Amplitude Limited Signal," Typescript, April 8, 1954, Bell Laboratories, 6 pp. + 1 Fig.

File 78:

[95] "Concavity of Transmission Rate as a Function of Input Probabilities," Memorandum MM 55-114-28, June 8, 1955, Bell Laboratories.

File 104:

[104] "Information Theory," Seminar Notes, Massachusetts Institute of Technology, 1956 and succeeding years. Contains the following sections:

"A skeleton key to the information theory notes," 3 pp. "Bounds on the tails of martingales and related questions," 19 pp. "Some useful inequalities for distribution functions," 3 pp. "A lower bound on the tail of a distribution," 9 pp. "A combinatorial theorem," 1 p. "Some results on determinants," 3 pp. "Upper and lower bounds for powers of a matrix with non-negative elements," 3 pp. "The number of sequences of a given length," 3 pp. "Characteristic for a language with independent letters," 4 pp. "The probability of error in optimal codes," 5 pp. "Zero error codes and the zero error capacity C_0 ," 10 pp. "Lower bound for P_{ef} for a completely connected channel with feedback," 1 p. "A lower bound for P_e when $R > C$," 2 pp. "A lower bound for P_e ," 2 pp. "Lower bound with one type of input and many types of output," 3 pp. "Application of 'sphere-packing' bounds to feedback case," 8 pp. "A result for the memoryless feedback channel," 1 p. "Continuity of $P_{e\ opt}$ as a function of transition probabilities," 1 p. "Codes of a fixed composition," 1 p. "Relation of P_e to ρ ," 2 pp. "Bound on P_e for random code by simple threshold argument," 4 pp. "A bound on P_e for a random code," 3 pp. "The Feinstein bound," 2 pp. "Relations between probability and minimum word separation," 4 pp.

File 104:

"Inequalities for decodable codes," 3 pp. "Convexity of channel capacity as a function of transition probabilities," 1 pp. "A geometric interpretation of channel capacity," 6 pp. "Log moment generating function for the square of a Gaussian variate," 2 pp. "Upper bound on P_e for Gaussian channel by expurgated random code," 2 pp. "Lower bound on P_e in Gaussian channel by minimum distance argument," 2 pp. "The sphere packing bound for the Gaussian power limited channel," 4 pp. "The T -terminal channel," 7 pp. "Conditions for constant mutual information," 2 pp. "The central limit theorem with large deviations," 6 pp. "The Chernoff inequality," 2 pp. "Upper and lower bounds on the tails of distributions," 4 pp. "Asymptotic behavior of the distribution function," 5 pp. "Generalized Chebycheff and Chernoff inequalities," 1 p. "Channels with side information at the transmitter," 13 pp. "Some miscellaneous results in coding theory," 15 pp. "Error probability bounds for noisy channels," 20 pp.

File 105:

[105] "Reliable Machines from Unreliable Components," notes of five lectures, Massachusetts Institute of Technology, Spring 1956, 24 pp.

File 105:

[106] "The Portfolio Problem, and How to Pay the Forecaster," lecture notes taken by W. W. Peterson, Massachusetts Institute of Technology, Spring, 1956, 8 pp.

File 105:

[107] "Notes on Relation of Error Probability to Delay in a Noisy Channel," notes of a lecture, Massachusetts Institute of Technology, Aug. 30, 1956, 3 pp.

File 105:

[108] "Notes on the Kelly Betting Theory of Noisy Information," notes of a lecture, Massachusetts Institute of Technology, Aug. 31, 1956, 2 pp.

File 105:

[124] "The Fourth-Dimensional Twist, or a Modest Proposal in Aid of the American Driver in England," typescript, All Souls College, Oxford, Trinity term, 1978, 7 pp. + 8 figs.

File 105:

[127] "A Rubric on Rubik Cubics," Typescript, circa 1982, 6 pp.

NJA Sloane

Claude Elwood Shannon
Miscellaneous Writings

Edited by

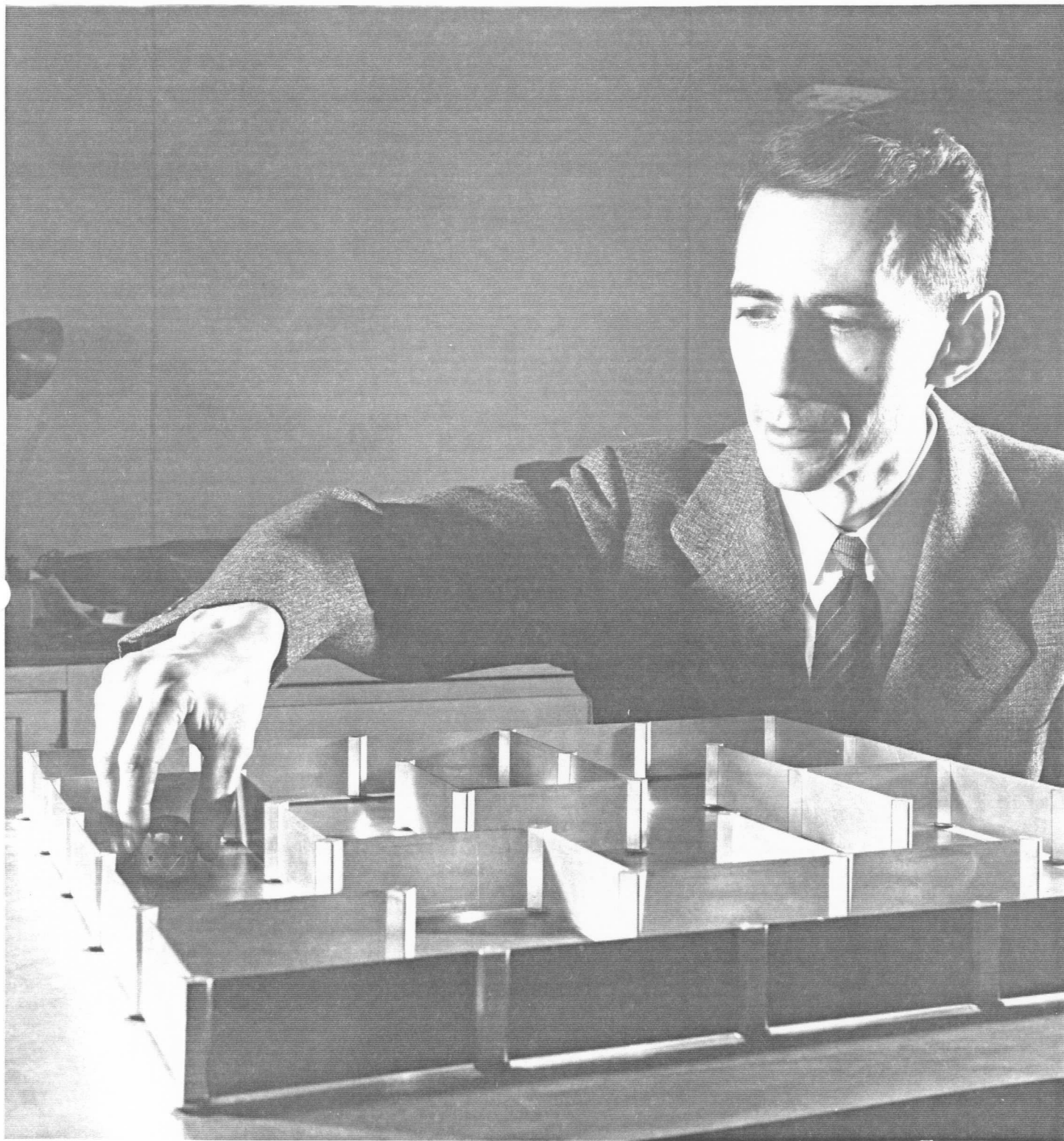
N. J. A. Sloane
Aaron D. Wyner

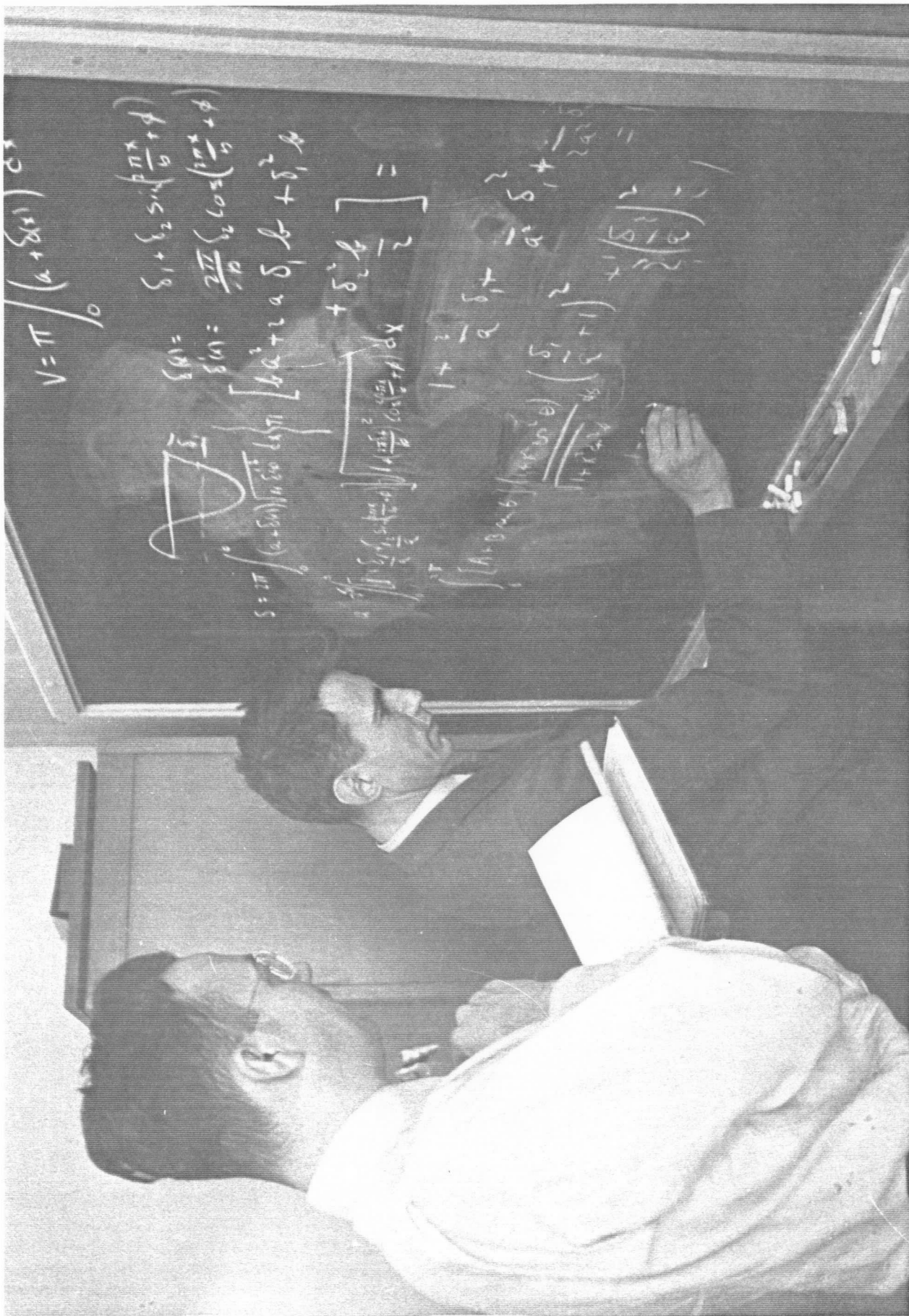
Mathematical Sciences Research Center, AT&T Bell Laboratories, Murray Hill,
New Jersey 07974

Preface

This volume contains all of Claude Elwood Shannon's writings that we did not include in his *Collected Papers*.*

* Claude Elwood Shannon: *Collected Papers*, edited by N. J. A. Sloane and A. D. Wyner, IEEE Press, New York, 1993, xlv + 924 pp. ISBN 0-7803-0434-9.





$$V = \pi \int_0^a (a + \delta(x))^2 dx$$

$$\delta(x) = \delta_1 + \delta_2 \sin\left(\frac{2\pi x}{b} + \phi\right)$$

$$\delta'(x) = \frac{2\pi}{b} \delta_2 \cos\left(\frac{2\pi x}{b} + \phi\right)$$

$$\left[b\delta_1^2 + 2a\delta_1\delta_2 + \delta_2^2 \right] =$$



$$S = 2\pi \int_0^a (a + \delta(x)) \sqrt{1 + \delta'(x)^2} dx$$

$$\int_0^a \left[1 + \frac{2}{a} \delta_1 + \frac{\delta_2^2}{a^2} + \frac{2\delta_1\delta_2}{a} \cos\left(\frac{2\pi x}{b} + \phi\right) \right] dx$$

$$\int_0^a \left[1 + \frac{2}{a} \delta_1 + \frac{\delta_2^2}{a^2} + \frac{2\delta_1\delta_2}{a} \cos\left(\frac{2\pi x}{b} + \phi\right) \right] dx$$

Contents

Photograph of Claude Shannon at Bell Labs in May 1952. Caption: "In 1952, Claude E. Shannon of Bell Laboratories devised an experiment to illustrate the capabilities of telephone relays. Here, an electrical mouse finds its way unerringly through a maze, guided by information remembered in the kind of switching relays used in dial telephone systems. Experiments with the mouse helped stimulate Bell Laboratories researchers to think of new ways to use the logical powers of computers for operations other than numerical calculation."

Photograph of Claude Shannon and Dave Hagelbarger at Bell Labs in March 1955. Caption: "Claude Shannon, the originator of Information Theory, at the board and Dave Hagelbarger work out some equations needed. Their current projects include work on automata-advanced type of computing machines which are able to perform various thought functions."

Photograph of Claude Shannon taken in 1980's. Photographer unknown.

Preface

Bibliography of Claude Elwood Shannon. Comments such as "Included in Part B" refer to Parts A, B, C, D of the *Collected Papers* mentioned in the Preface.

This volume contains the following items. Bracketed numbers refer to the bibliography.

- [5] "The Use of the Lakatos-Hickman Relay in a Subscriber Sender," Memorandum MM 40-130-179, August 3, 1940, Bell Laboratories, 7 pp. + 8 figs.
- [7] "A Study of the Deflection Mechanism and Some Results on Rate Finders," Report to National Defense Research Committee, Div. 7-311-M1, circa April, 1941, 37 pp. + 15 figs.
- [9] "A Height Data Smoothing Mechanism," Report to National Defense Research Committee, Div. 7-313.2-M1, Princeton Univ., May 26, 1941, 9 pp. + 9 figs.
- [11] "Some Experimental Results on the Deflection Mechanism," Report to National Defense Research Committee, Div. 7-311-M1, June 26, 1941, 11 pp.
- [12] "Criteria for Consistency and Uniqueness in Relay Circuits," Typescript, Sept. 8, 1941, 5 pp. + 3 figs.
- [16] (With W. Feller) "On the Integration of the Ballistic Equations on the Aberdeen Analyzer," Applied Mathematics Panel Report No. 28.1, National Defense Research Committee, July 15, 1943, 9 pp.
- [19] "Two New Circuits for Alternate Pulse Counting," Typescript, May 29, 1944, Bell Laboratories, 2 pp. + 3 Figs.

- [20] "Counting Up or Down With Pulse Counters," Typescript, May 31, 1944, Bell Laboratories, 1 p. + 1 fig.
- [21] (With B. M. Oliver) "Circuits for a P.C.M. Transmitter and Receiver," Memorandum MM 44-110-37, June 1, 1944, Bell Laboratories, 4 pp., 11 figs.
- [23] "Pulse Shape to Minimize Bandwidth With Nonoverlapping Pulses," Typescript, August 4, 1944, Bell Laboratories, 4 pp.
- [24] "A Mathematical Theory of Cryptography," Memorandum MM 45-110-02, Sept. 1, 1945, Bell Laboratories, 114 pp. + 25 figs.
- [26] "Mixed Statistical Determinate Systems," Typescript, Sept. 19, 1945, Bell Laboratories, 17 pp.
- [27] (With R. B. Blackman and H. W. Bode) "Data Smoothing and Prediction in Fire-Control Systems," Summary Technical Report, Div. 7, National Defense Research Committee, Vol. 1, *Gunfire Control*, Washington, DC, 1946, pp. 71-159 and 166-167. AD 200795. Also in National Military Establishment Research and Development Board, Report #13 MGC 12/1, August 15, 1948. Superseded by [51] and by R. B. Blackman, *Linear Data-Smoothing and Prediction in Theory and Practice*, Addison-Wesley, Reading, Mass., 1965.
- [30] (With C. L. Dolph) "The Transient Behavior of a Large Number of Four-Terminal Unilateral Linear Networks Connected in Tandem," Memorandum MM 46-110-49, April 10, 1946, Bell Laboratories, 34 pp. + 16 figs.
- [31] "Electronic Methods in Telephone Switching," Typescript, October 17, 1946, Bell Laboratories, 5 pp. + 1 fig.
- [32] "Some Generalizations of the Sampling Theorem," Typescript, March 4, 1948, 5 pp. + 1 fig.
- [34] "The Normal Ergodic Ensembles of Functions," Typescript, March 15, 1948, 5 pp.
- [35] "Systems Which Approach the Ideal as $P/N \rightarrow \infty$," Typescript, March 15, 1948, 2 pp.
- [36] "Theorems on Statistical Sequences," Typescript, March 15, 1948, 8 pp.
- [45] "Significance and Application [of Communication Research]," *Symposium on Communication Research, 11-13 October, 1948*, Research and Development Board, Department of Defense, Washington, DC, pp. 14-23, 1948.
- [46] "Note on Certain Transcendental Numbers," Typescript, October 27, 1948, Bell Laboratories, 1 p.
- [47] "A Case of Efficient Coding for a Very Noisy Channel," Typescript, Nov. 18, 1948, Bell Laboratories, 2 pp.
- [48] "Note on Reversing a Discrete Markhoff Process," Typescript, Dec. 6 1948, Bell Laboratories, 2 pp. + 2 Figs.

- [49] "Information Theory," Typescript of abstract of talk for American Statistical Society, 1949, 5 pp.
- [58] "Proof of an Integration Formula," Typescript, circa 1950, Bell Laboratories, 2 pp.
- [59] "A Digital Method of Transmitting Information," Typescript, no date, circa 1950, Bell Laboratories, 3 pp.
- [72] "Creative Thinking," Typescript, March 20, 1952, Bell Laboratories, 10 pp.
- [74] (With E. F. Moore) "The Relay Circuit Analyzer," Memorandum MM 53-1400-9, March 31, 1953, Bell Laboratories, 14 pp. + 4 figs.
- [77] "Throbac - Circuit Operation," Typescript, April 9, 1953, Bell Laboratories, 7 pp.
- [78] "Tower of Hanoi," Typescript, April 20, 1953, Bell Laboratories, 4 pp.
- [81] "Mathmanship or How to Give an Explicit Solution Without Actually Solving the Problem," Typescript, June 3, 1953, Bell Laboratories, 2 pp.
- [84] (With E. F. Moore) "The Relay Circuit Synthesizer," Memorandum MM 53-140-52, November 30, 1953, Bell Laboratories, 22 pp. + 5 figs.
- [87] "Bounds on the Derivatives and Rise Time of a Band and Amplitude Limited Signal," Typescript, April 8, 1954, Bell Laboratories, 6 pp. + 1 Fig.
- [95] "Concavity of Transmission Rate as a Function of Input Probabilities," Memorandum MM 55-114-28, June 8, 1955, Bell Laboratories.
- [104] "Information Theory," Seminar Notes, Massachusetts Institute of Technology, 1956 and succeeding years. Contains the following sections:

"A skeleton key to the information theory notes," 3 pp. "Bounds on the tails of martingales and related questions," 19 pp. "Some useful inequalities for distribution functions," 3 pp. "A lower bound on the tail of a distribution," 9 pp. "A combinatorial theorem," 1 p. "Some results on determinants," 3 pp. "Upper and lower bounds for powers of a matrix with non-negative elements," 3 pp. "The number of sequences of a given length," 3 pp. "Characteristic for a language with independent letters," 4 pp. "The probability of error in optimal codes," 5 pp. "Zero error codes and the zero error capacity C_0 ," 10 pp. "Lower bound for P_{ef} for a completely connected channel with feedback," 1 p. "A lower bound for P_e when $R > C$," 2 pp. "A lower bound for P_e ," 2 pp. "Lower bound with one type of input and many types of output," 3 pp. "Application of 'sphere-packing' bounds to feedback case," 8 pp. "A result for the memoryless feedback channel," 1 p. "Continuity of $P_{e\ opt}$ as a function of transition probabilities," 1 p. "Codes of a fixed composition," 1 p. "Relation of P_e to ρ ," 2 pp. "Bound on P_e for random code by simple threshold argument," 4 pp. "A bound on P_e for a random code," 3 pp. "The Feinstein bound," 2 pp. "Relations between probability and minimum word separation," 4 pp.

- "Inequalities for decodable codes," 3 pp. "Convexity of channel capacity as a function of transition probabilities," 1 pp. "A geometric interpretation of channel capacity," 6 pp. "Log moment generating function for the square of a Gaussian variate," 2 pp. "Upper bound on P_e for Gaussian channel by expurgated random code," 2 pp. "Lower bound on P_e in Gaussian channel by minimum distance argument," 2 pp. "The sphere packing bound for the Gaussian power limited channel," 4 pp. "The T -terminal channel," 7 pp. "Conditions for constant mutual information," 2 pp. "The central limit theorem with large deviations," 6 pp. "The Chernoff inequality," 2 pp. "Upper and lower bounds on the tails of distributions," 4 pp. "Asymptotic behavior of the distribution function," 5 pp. "Generalized Chebycheff and Chernoff inequalities," 1 p. "Channels with side information at the transmitter," 13 pp. "Some miscellaneous results in coding theory," 15 pp. "Error probability bounds for noisy channels," 20 pp.
- [105] "Reliable Machines from Unreliable Components," notes of five lectures, Massachusetts Institute of Technology, Spring 1956, 24 pp.
- [106] "The Portfolio Problem, and How to Pay the Forecaster," lecture notes taken by W. W. Peterson, Massachusetts Institute of Technology, Spring, 1956, 8 pp.
- [107] "Notes on Relation of Error Probability to Delay in a Noisy Channel," notes of a lecture, Massachusetts Institute of Technology, Aug. 30, 1956, 3 pp.
- [108] "Notes on the Kelly Betting Theory of Noisy Information," notes of a lecture, Massachusetts Institute of Technology, Aug. 31, 1956, 2 pp.
- [124] "The Fourth-Dimensional Twist, or a Modest Proposal in Aid of the American Driver in England," typescript, All Souls College, Oxford, Trinity term, 1978, 7 pp. + 8 figs.
- [127] "A Rubric on Rubik Cubics," Typescript, circa 1982, 6 pp.

Bibliography of Claude Elwood Shannon

- [1] "A Symbolic Analysis of Relay and Switching Circuits," *Transactions American Institute of Electrical Engineers*, Vol. 57 (1938), pp. 713-723. (Received March 1, 1938.) Included in Part B.
- [2] Letter to Vannevar Bush, Feb. 16, 1939. Printed in F.-W. Hagemeyer, *Die Entstehung von Informationskonzepten in der Nachrichtentechnik: eine Fallstudie zur Theoriebildung in der Technik in Industrie- und Kriegsforschung* [The Origin of Information Theory Concepts in Communication Technology: Case Study for Engineering Theory-Building in Industrial and Military Research], Doctoral Dissertation, Free Univ. Berlin, Nov. 8, 1979, 570 pp. Included in Part A.
- [3] "An Algebra for Theoretical Genetics," Ph.D. Dissertation, Department of Mathematics, Massachusetts Institute of Technology, April 15, 1940, 69 pp. Included in Part C.
- [4] "A Theorem on Color Coding," Memorandum 40-130-153, July 8, 1940, Bell Laboratories. Superseded by "A Theorem on Coloring the Lines of a Network." Not included.
- [5] "The Use of the Lakatos-Hickman Relay in a Subscriber Sender," Memorandum MM 40-130-179, August 3, 1940, Bell Laboratories, 7 pp. + 8 figs. Included in this volume.
- [6] "Mathematical Theory of the Differential Analyzer," *Journal of Mathematics and Physics*, Vol. 20 (1941), pp. 337-354. Included in Part B.
- [7] "A Study of the Deflection Mechanism and Some Results on Rate Finders," Report to National Defense Research Committee, Div. 7-311-M1, circa April, 1941, 37 pp. + 15 figs. Included in this volume.
- [8] "Backlash in Overdamped Systems," Report to National Defense Research Committee, Princeton Univ., May 14, 1941, 6 pp. Abstract only included in Part B.
- [9] "A Height Data Smoothing Mechanism," Report to National Defense Research Committee, Div. 7-313.2-M1, Princeton Univ., May 26, 1941, 9 pp. + 9 figs. Included in this volume.
- [10] "The Theory of Linear Differential and Smoothing Operators," Report to National Defense Research Committee, Div. 7-313.1-M1, Princeton Univ., June 8, 1941, 11 pp. Not included.
- [11] "Some Experimental Results on the Deflection Mechanism," Report to National Defense Research Committee, Div. 7-311-M1, June 26, 1941, 11 pp. Included in this volume.

- [12] "Criteria for Consistency and Uniqueness in Relay Circuits," Typescript, Sept. 8, 1941, 5 pp. + 3 figs. Included in this volume.
- [13] "The Theory and Design of Linear Differential Equation Machines," Report to the Services 20, Div. 7-311-M2, Jan. 1942, Bell Laboratories, 73 pp. + 30 figs. Included in Part B.
- [14] (With John Riordan) "The Number of Two-Terminal Series-Parallel Networks," *Journal of Mathematics and Physics*, Vol. 21 (August, 1942), pp. 83-93. Included in Part B.
- [15] "Analogue of the Vernam System for Continuous Time Series," Memorandum MM 43-110-44, May 10, 1943, Bell Laboratories, 4 pp. + 4 figs. Included in Part A.
- [16] (With W. Feller) "On the Integration of the Ballistic Equations on the Aberdeen Analyzer," Applied Mathematics Panel Report No. 28.1, National Defense Research Committee, July 15, 1943, 9 pp. Included in this volume.
- [17] "Pulse Code Modulation," Memorandum MM 43-110-43, December 1, 1943, Bell Laboratories. Not included.
- [18] "Feedback Systems with Periodic Loop Closure," Memorandum MM 44-110-32, March 16, 1944, Bell Laboratories. Not included.
- [19] "Two New Circuits for Alternate Pulse Counting," Typescript, May 29, 1944, Bell Laboratories, 2 pp. + 3 Figs. Included in this volume.
- [20] "Counting Up or Down With Pulse Counters," Typescript, May 31, 1944, Bell Laboratories, 1 p. + 1 fig. Included in this volume.
- [21] (With B. M. Oliver) "Circuits for a P.C.M. Transmitter and Receiver," Memorandum MM 44-110-37, June 1, 1944, Bell Laboratories, 4 pp., 11 figs. Included in this volume.
- [22] "The Best Detection of Pulses," Memorandum MM 44-110-28, June 22, 1944, Bell Laboratories, 3 pp. Included in Part A.
- [23] "Pulse Shape to Minimize Bandwidth With Nonoverlapping Pulses," Typescript, August 4, 1944, Bell Laboratories, 4 pp. Included in this volume.
- [24] "A Mathematical Theory of Cryptography," Memorandum MM 45-110-02, Sept. 1, 1945, Bell Laboratories, 114 pp. + 25 figs. Superseded by the following paper. Included in this volume.
- [25] "Communication Theory of Secrecy Systems," *Bell System Technical Journal*, Vol. 28 (1949), pp. 656-715. "The material in this paper appeared originally in a confidential report 'A Mathematical Theory of Cryptography', dated Sept. 1, 1945, which has now been declassified." Included in Part A.

- [26] "Mixed Statistical Determinate Systems," Typescript, Sept. 19, 1945, Bell Laboratories, 17 pp. Included in this volume.
- [27] (With R. B. Blackman and H. W. Bode) "Data Smoothing and Prediction in Fire-Control Systems," Summary Technical Report, Div. 7, National Defense Research Committee, Vol. 1, *Gunfire Control*, Washington, DC, 1946, pp. 71-159 and 166-167. AD 200795. Also in National Military Establishment Research and Development Board, Report #13 MGC 12/1, August 15, 1948. Superseded by [51] and by R. B. Blackman, *Linear Data-Smoothing and Prediction in Theory and Practice*, Addison-Wesley, Reading, Mass., 1965. Included in this volume.
- [28] (With B. M. Oliver) "Communication System Employing Pulse Code Modulation," Patent 2,801,281. Filed Feb. 21, 1946, granted July 30, 1957. Not included.
- [29] (With B. D. Holbrook) "A Sender Circuit For Panel or Crossbar Telephone Systems," Patent application circa 1946, application dropped April 13, 1948. Not included.
- [30] (With C. L. Dolph) "The Transient Behavior of a Large Number of Four-Terminal Unilateral Linear Networks Connected in Tandem," Memorandum MM 46-110-49, April 10, 1946, Bell Laboratories, 34 pp. + 16 figs. Included in this volume.
- [31] "Electronic Methods in Telephone Switching," Typescript, October 17, 1946, Bell Laboratories, 5 pp. + 1 fig. Included in this volume.
- [32] "Some Generalizations of the Sampling Theorem," Typescript, March 4, 1948, 5 pp. + 1 fig. Included in this volume.
- [33] (With J. R. Pierce and J. W. Tukey) "Cathode-Ray Device," Patent 2,576,040. Filed March 10, 1948, granted Nov. 20, 1951. Not included.
- [34] "The Normal Ergodic Ensembles of Functions," Typescript, March 15, 1948, 5 pp. Included in this volume.
- [35] "Systems Which Approach the Ideal as $P/N \rightarrow \infty$," Typescript, March 15, 1948, 2 pp. Included in this volume.
- [36] "Theorems on Statistical Sequences," Typescript, March 15, 1948, 8 pp. Included in this volume.
- [37] "A Mathematical Theory of Communication," *Bell System Technical Journal*, Vol. 27 (July and October 1948), pp. 379-423 and 623-656. Reprinted in D. Slepian, editor, *Key Papers in the Development of Information Theory*, IEEE Press, NY, 1974. Included in Part A.
- [38] (With Warren Weaver) *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, IL, 1949, vi + 117 pp. Reprinted (and repaginated) 1963. The section by Shannon is essentially identical

to the previous item. Not included.

- [39] (With Warren Weaver) *Mathematische Grundlagen der Informationstheorie*, Scientia Nova, Oldenbourg Verlag, München, 1976, pp. 143. German translation of the preceding book. Not included.
- [40] (With B. M. Oliver and J. R. Pierce) "The Philosophy of PCM," *Proceedings Institute of Radio Engineers*, Vol. 36 (1948), pp. 1324-1331. (Received May 24, 1948.) Included in Part A.
- [41] "Samples of Statistical English," Typescript, June 11, 1948, Bell Laboratories, 3 pp. Included in this volume.
- [42] "Network Rings," Typescript, June 11, 1948, Bell Laboratories, 26 pp. + 4 figs. Included in Part B.
- [43] "Communication in the Presence of Noise," *Proceedings Institute of Radio Engineers*, Vol. 37 (1949), pp. 10-21. (Received July 23, 1940 [1948?].) Reprinted in D. Slepian, editor, *Key Papers in the Development of Information Theory*, IEEE Press, NY, 1974. Reprinted in *Proceedings Institute of Electrical and Electronic Engineers*, Vol. 72 (1984), pp. 1192-1201. Included in Part A.
- [44] "A Theorem on Coloring the Lines of a Network," *Journal of Mathematics and Physics*, Vol. 28 (1949), pp. 148-151. (Received Sept. 14, 1948.) Included in Part B.
- [45] "Significance and Application [of Communication Research]," *Symposium on Communication Research, 11-13 October, 1948*, Research and Development Board, Department of Defense, Washington, DC, pp. 14-23, 1948. Included in this volume.
- [46] "Note on Certain Transcendental Numbers," Typescript, October 27, 1948, Bell Laboratories, 1 p. Included in this volume.
- [47] "A Case of Efficient Coding for a Very Noisy Channel," Typescript, Nov. 18, 1948, Bell Laboratories, 2 pp. Included in this volume.
- [48] "Note on Reversing a Discrete Markhoff Process," Typescript, Dec. 6 1948, Bell Laboratories, 2 pp. + 2 Figs. Included in this volume.
- [49] "Information Theory," Typescript of abstract of talk for American Statistical Society, 1949, 5 pp. Included in this volume.
- [50] "The Synthesis of Two-Terminal Switching Circuits," *Bell System Technical Journal*, Vol. 28 (Jan., 1949), pp. 59-98. Included in Part B.
- [51] (With H. W. Bode) "A Simplified Derivation of Linear Least Squares Smoothing and Prediction Theory," *Proceedings Institute of Radio Engineers*, Vol. 38 (1950), pp. 417-425. (Received July 13, 1949.) Included in Part B.

- [52] "Review of *Transformations on Lattices and Structures of Logic* by Stephen A. Kiss," *Proceedings Institute of Radio Engineers*, Vol. 37 (1949), p. 1163. Included in Part B.
- [53] "Review of *Cybernetics, or Control and Communication in the Animal and the Machine* by Norbert Wiener," *Proceedings Institute of Radio Engineers*, Vol. 37 (1949), p. 1305. Included in Part B.
- [54] "Programming a Computer for Playing Chess," *Philosophical Magazine*, Series 7, Vol. 41 (No. 314, March 1950), pp. 256-275. (Received Nov. 8, 1949.) Reprinted in D. N. L. Levy, editor, *Computer Chess Compendium*, Springer-Verlag, NY, 1988. Included in Part B.
- [55] "A Chess-Playing Machine," *Scientific American*, Vol. 182 (No. 2, February 1950), pp. 48-51. Reprinted in *The World of Mathematics*, edited by James R. Newman, Simon and Schuster, NY, Vol. 4, 1956, pp. 2124-2133. Included in Part B.
- [56] "Memory Requirements in a Telephone Exchange," *Bell System Technical Journal*, Vol. 29 (1950), pp. 343-349. (Received Dec. 7, 1949.) Included in Part B.
- [57] "A Symmetrical Notation for Numbers," *American Mathematical Monthly*, Vol. 57 (Feb., 1950), pp. 90-93. Included in Part B.
- [58] "Proof of an Integration Formula," Typescript, circa 1950, Bell Laboratories, 2 pp. Included in this volume.
- [59] "A Digital Method of Transmitting Information," Typescript, no date, circa 1950, Bell Laboratories, 3 pp. Included in this volume.
- [60] "Communication Theory — Exposition of Fundamentals," in "Report of Proceedings, Symposium on Information Theory, London, Sept., 1950," *Institute of Radio Engineers, Transactions on Information Theory*, No. 1 (February, 1953), pp. 44-47. Included in Part A.
- [61] "General Treatment of the Problem of Coding," in "Report of Proceedings, Symposium on Information Theory, London, Sept., 1950," *Institute of Radio Engineers, Transactions on Information Theory*, No. 1 (February, 1953), pp. 102-104. Included in Part A.
- [62] "The Lattice Theory of Information," in "Report of Proceedings, Symposium on Information Theory, London, Sept., 1950," *Institute of Radio Engineers, Transactions on Information Theory*, No. 1 (February, 1953), pp. 105-107. Included in Part A.
- [63] (With E. C. Cherry, S. H. Moss, Dr. Uttley, I. J. Good, W. Lawrence and W. P. Anderson) "Discussion of Preceding Three Papers," in "Report of Proceedings, Symposium on Information Theory, London, Sept., 1950," *Institute of Radio Engineers, Transactions on Information Theory*, No. 1 (February, 1953), pp. 169-174. Included in Part A.

- [64] "Review of *Description of a Relay Computer*, by the Staff of the [Harvard] Computation Laboratory," *Proceedings Institute of Radio Engineers*, Vol. 38 (1950), p. 449. Included in Part B.
- [65] "Recent Developments in Communication Theory," *Electronics*, Vol. 23 (April, 1950), pp. 80-83. Included in Part A.
- [66] German translation of [65], in *Tech. Mitt. P.T.T.*, Bern, Vol. 28 (1950), pp. 337-342. Not included.
- [67] "A Method of Power or Signal Transmission To a Moving Vehicle," Memorandum for Record, July 19, 1950, Bell Laboratories, 2 pp. + 4 figs. Included in Part B.
- [68] "Some Topics in Information Theory," in *Proceedings International Congress of Mathematicians (Cambridge, Mass., Aug. 30 - Sept. 6, 1950)*, American Mathematical Society, Vol. II (1952), pp. 262-263. Included in Part A.
- [69] "Prediction and Entropy of Printed English," *Bell System Technical Journal*, Vol. 30 (1951), pp. 50-64. (Received Sept. 15, 1950.) Reprinted in D. Slepian, editor, *Key Papers in the Development of Information Theory*, IEEE Press, NY, 1974. Included in Part A.
- [70] "Presentation of a Maze Solving Machine," in *Cybernetics: Circular, Causal and Feedback Mechanisms in Biological and Social Systems, Transactions Eighth Conference, March 15-16, 1951, New York, N. Y.*, edited by H. von Foerster, M. Mead and H. L. Teuber, Josiah Macy Jr. Foundation, New York, 1952, pp. 169-181. Included in Part B.
- [71] "Control Apparatus," Patent application Aug. 1951, dropped Jan. 21, 1954. Not included.
- [72] "Creative Thinking," Typescript, March 20, 1952, Bell Laboratories, 10 pp. Included in this volume.
- [73] "A Mind-Reading (?) Machine," Typescript, March 18, 1953, Bell Laboratories, 4 pp. Included in Part B.
- [74] (With E. F. Moore) "The Relay Circuit Analyzer," Memorandum MM 53-1400-9, March 31, 1953, Bell Laboratories, 14 pp. + 4 figs. Included in this volume.
- [75] "The Potentialities of Computers," Typescript, April 3, 1953, Bell Laboratories. Included in Part B.
- [76] "Throbac I," Typescript, April 9, 1953, Bell Laboratories, 5 pp. Included in Part B.
- [77] "Throbac - Circuit Operation," Typescript, April 9, 1953, Bell Laboratories, 7 pp. Included in this volume.

- [78] "Tower of Hanoi," Typescript, April 20, 1953, Bell Laboratories, 4 pp. Included in this volume.
- [79] (With E. F. Moore) "Electrical Circuit Analyzer," Patent 2,776,405. Filed May 18, 1953, granted Jan. 1, 1957. Not included.
- [80] (With E. F. Moore) "Machine Aid for Switching Circuit Design," *Proceedings Institute of Radio Engineers*, Vol. 41 (1953), pp. 1348-1351. (Received May 28, 1953.) Included in Part B.
- [81] "Mathmanship or How to Give an Explicit Solution Without Actually Solving the Problem," Typescript, June 3, 1953, Bell Laboratories, 2 pp. Included in this volume.
- [82] "Computers and Automata," *Proceedings Institute of Radio Engineers*, Vol. 41 (1953), pp. 1234-1241. (Received July 17, 1953.) Reprinted in *Methodos*, Vol. 6 (1954), pp. 115-130. Included in Part B.
- [83] "Realization of All 16 Switching Functions of Two Variables Requires 18 Contacts," Memorandum MM 53-1400-40, November 17, 1953, Bell Laboratories, 4 pp. + 2 figs. Included in Part B.
- [84] (With E. F. Moore) "The Relay Circuit Synthesizer," Memorandum MM 53-140-52, November 30, 1953, Bell Laboratories, 26 pp. + 5 figs. Included in this volume.
- [85] (With D. W. Hagelbarger) "A Relay Laboratory Outfit for Colleges," Memorandum MM 54-114-17, January 10, 1954, Bell Laboratories. Included in Part B.
- [86] "Efficient Coding of a Binary Source With One Very Infrequent Symbol," Memorandum MM 54-114-7, January 29, 1954, Bell Laboratories. Included in Part A.
- [87] "Bounds on the Derivatives and Rise Time of a Band and Amplitude Limited Signal," Typescript, April 8, 1954, Bell Laboratories, 6 pp. + 1 Fig. Included in this volume.
- [88] (With Edward F. Moore) "Reliable Circuits Using Crummy Relays," Memorandum 54-114-42, Nov. 29, 1954, Bell Laboratories. Published as the following two items.
- [89] (With Edward F. Moore) "Reliable Circuits Using Less Reliable Relays I," *Journal Franklin Institute*, Vol. 262 (Sept., 1956), pp. 191-208. Included in Part B.
- [90] (With Edward F. Moore) "Reliable Circuits Using Less Reliable Relays II," *Journal Franklin Institute*, Vol. 262 (Oct., 1956), pp. 281-297. Included in Part B.
- [91] (Edited jointly with John McCarthy) *Automata Studies*, Annals of Mathematics Studies Number 34, Princeton University Press, Princeton,

- NJ, 1956, ix + 285 pp. The Preface, Table of Contents, and the two papers by Shannon are included in Part B.
- [92] (With John McCarthy), *Studien zur Theorie der Automaten*, München, 1974. (German translation of the preceding work.)
- [93] "A Universal Turing Machine With Two Internal States," Memorandum 54-114-38, May 15, 1954, Bell Laboratories. Published in *Automata Studies*, pp. 157-165. Included in Part B.
- [94] (With Karel de Leeuw, Edward F. Moore and N. Shapiro) "Computability by Probabilistic Machines," Memorandum 54-114-37, Oct. 21, 1954, Bell Laboratories. Published in [87], pp. 183-212. Included in Part B.
- [95] "Concavity of Transmission Rate as a Function of Input Probabilities," Memorandum MM 55-114-28, June 8, 1955, Bell Laboratories. Included in this volume.
- [96] "Some Results on Ideal Rectifier Circuits," Memorandum MM 55-114-29, June 8, 1955, Bell Laboratories. Included in Part B.
- [97] "The Simultaneous Synthesis of s Switching Functions of n Variables," Memorandum MM 55-114-30, June 8, 1955, Bell Laboratories. Included in Part B.
- [98] (With D. W. Hagelbarger) "Concavity of Resistance Functions," *Journal Applied Physics*, Vol. 27 (1956), pp. 42-43. (Received August 1, 1955.) Included in Part B.
- [99] "Game Playing Machines," *Journal Franklin Institute*, Vol. 260 (1955), pp. 447-453. (Delivered Oct. 19, 1955.) Included in Part B.
- [100] "Information Theory," *Encyclopedia Britannica*, Chicago, IL, 14th Edition, 1968 printing, Vol. 12, pp. 246B-249. (Written circa 1955.) Included in Part A.
- [101] "Cybernetics," *Encyclopedia Britannica*, Chicago, IL, 14th Edition, 1968 printing, Vol. 12. (Written circa 1955.) Not included.
- [102] "The Rate of Approach to Ideal Coding (Abstract)," *Proceedings Institute of Radio Engineers*, Vol. 43 (1955), p. 356. Included in Part A.
- [103] "The Bandwagon (Editorial)," *Institute of Radio Engineers, Transactions on Information Theory*, Vol. IT-2 (March, 1956), p. 3. Included in Part A.
- [104] "Information Theory," Seminar Notes, Massachusetts Institute of Technology, 1956 and succeeding years. Included in this volume. Contains the following sections:
- "A skeleton key to the information theory notes," 3 pp. "Bounds on the

tails of martingales and related questions," 19 pp. "Some useful inequalities for distribution functions," 3 pp. "A lower bound on the tail of a distribution," 9 pp. "A combinatorial theorem," 1 p. "Some results on determinants," 3 pp. "Upper and lower bounds for powers of a matrix with non-negative elements," 3 pp. "The number of sequences of a given length," 3 pp. "Characteristic for a language with independent letters," 4 pp. "The probability of error in optimal codes," 5 pp. "Zero error codes and the zero error capacity C_0 ," 10 pp. "Lower bound for P_{ef} for a completely connected channel with feedback," 1 p. "A lower bound for P_e when $R > C$," 2 pp. "A lower bound for P_e ," 2 pp. "Lower bound with one type of input and many types of output," 3 pp. "Application of 'sphere-packing' bounds to feedback case," 8 pp. "A result for the memoryless feedback channel," 1 p. "Continuity of $P_{e\ opt}$ as a function of transition probabilities," 1 p. "Codes of a fixed composition," 1 p. "Relation of P_e to ρ ," 2 pp. "Bound on P_e for random code by simple threshold argument," 4 pp. "A bound on P_e for a random code," 3 pp. "The Feinstein bound," 2 pp. "Relations between probability and minimum word separation," 4 pp. "Inequalities for decodable codes," 3 pp. "Convexity of channel capacity as a function of transition probabilities," 1 pp. "A geometric interpretation of channel capacity," 6 pp. "Log moment generating function for the square of a Gaussian variate," 2 pp. "Upper bound on P_e for Gaussian channel by expurgated random code," 2 pp. "Lower bound on P_e in Gaussian channel by minimum distance argument," 2 pp. "The sphere packing bound for the Gaussian power limited channel," 4 pp. "The T -terminal channel," 7 pp. "Conditions for constant mutual information," 2 pp. "The central limit theorem with large deviations," 6 pp. "The Chernoff inequality," 2 pp. "Upper and lower bounds on the tails of distributions," 4 pp. "Asymptotic behavior of the distribution function," 5 pp. "Generalized Chebycheff and Chernoff inequalities," 1 p. "Channels with side information at the transmitter," 13 pp. "Some miscellaneous results in coding theory," 15 pp. "Error probability bounds for noisy channels," 20 pp.

- [105] "Reliable Machines from Unreliable Components," notes of five lectures, Massachusetts Institute of Technology, Spring 1956, 24 pp. Not included.
- [106] "The Portfolio Problem, and How to Pay the Forecaster," lecture notes taken by W. W. Peterson, Massachusetts Institute of Technology, Spring, 1956, 8 pp. Included in this volume.
- [107] "Notes on Relation of Error Probability to Delay in a Noisy Channel," notes of a lecture, Massachusetts Institute of Technology, Aug. 30, 1956, 3 pp. Included in this volume.
- [108] "Notes on the Kelly Betting Theory of Noisy Information," notes of a lecture, Massachusetts Institute of Technology, Aug. 31, 1956, 2 pp.

Included in this volume.

- [109] "The Zero Error Capacity of a Noisy Channel," *Institute of Radio Engineers, Transactions on Information Theory*, Vol. IT-2 (September, 1956), pp. S8-S19. Reprinted in D. Slepian, editor, *Key Papers in the Development of Information Theory*, IEEE Press, NY, 1974. Included in Part A.
- [110] (With Peter Elias and Amiel Feinstein) "A Note on the Maximum Flow Through a Network," *Institute of Radio Engineers, Transactions on Information Theory*, Vol. IT-2 (December, 1956), pp. 117-119. (Received July 11, 1956.) Included in Part B.
- [111] "Certain Results in Coding Theory for Noisy Channels," *Information and Control*, Vol. 1 (1957), pp. 6-25. (Received April 22, 1957.) Reprinted in D. Slepian, editor, *Key Papers in the Development of Information Theory*, IEEE Press, NY, 1974. Included in Part A.
- [112] "Geometrische Deutung einiger Ergebnisse bei der Berechnung der Kanal Kapazität" [Geometrical meaning of some results in the calculation of channel capacity], *Nachrichtentechnische Zeit. (N.T.Z.)*, Vol. 10 (No. 1, January 1957), pp. 1-4. Not included, since the English version is included.
- [113] "Some Geometrical Results in Channel Capacity," *Verband Deutsche Elektrotechniker Fachber.*, Vol. 19 (II) (1956), pp. 13-15 = *Nachrichtentechnische Fachber. (N.T.F.)*, Vol. 6 (1957). English version of the preceding work. Included in Part A.
- [114] "Von Neumann's Contribution to Automata Theory," *Bulletin American Mathematical Society*, Vol. 64 (No. 3, Part 2, 1958), pp. 123-129. (Received Feb. 10, 1958.) Included in Part B.
- [115] "A Note on a Partial Ordering for Communication Channels," *Information and Control*, Vol. 1 (1958), pp. 390-397. (Received March 24, 1958.) Reprinted in D. Slepian, editor, *Key Papers in the Development of Information Theory*, IEEE Press, NY, 1974. Included in Part A.
- [116] "Channels With Side Information at the Transmitter," *IBM Journal Research and Development*, Vol. 2 (1958), pp. 289-293. (Received Sept. 15, 1958.) Reprinted in D. Slepian, editor, *Key Papers in the Development of Information Theory*, IEEE Press, NY, 1974. Included in Part A.
- [117] "Probability of Error for Optimal Codes in a Gaussian Channel," *Bell System Technical Journal*, Vol. 38 (1959), pp. 611-656. (Received Oct. 17, 1958.) Included in Part A.
- [118] "Coding Theorems for a Discrete Source With a Fidelity Criterion," *Institute of Radio Engineers, International Convention Record*, Vol. 7

- (Part 4, 1959), pp. 142-163. Reprinted with changes in *Information and Decision Processes*, edited by R. E. Machol, McGraw-Hill, NY, 1960, pp. 93-126. Reprinted in D. Slepian, editor, *Key Papers in the Development of Information Theory*, IEEE Press, NY, 1974. Included in Part A.
- [119] "Two-Way Communication Channels," in *Proceedings Fourth Berkeley Symposium Probability and Statistics, June 20 - July 30, 1960*, edited by J. Neyman, Univ. Calif. Press, Berkeley, CA, Vol. 1, 1961, pp. 611-644. Reprinted in D. Slepian, editor, *Key Papers in the Development of Information Theory*, IEEE Press, NY, 1974. Included in Part A.
- [120] "Computers and Automation — Progress and Promise in the Twentieth Century," *Man, Science, Learning and Education. The Semicentennial Lectures at Rice University*, edited by S. W. Higginbotham, Supplement 2 to Vol. XLIX, Rice University Studies, Rice Univ., 1963, pp. 201-211. Included in Part B.
- [121] *Papers in Information Theory and Cybernetics* (in Russian), Izd. Inostr. Lit., Moscow, 1963, 824 pp. Edited by R. L. Dobrushin and O. B. Lupanova, preface by A. N. Kolmogorov. Contains Russian translations of [1], [6], [14], [25], [37], [40], [43], [44], [50], [51], [54]-[56], [65], [68]-[70], [80], [82], [89], [90], [93], [94], [99], [103], [109]-[111], [113]-[119].
- [122] (With R. G. Gallager and E. R. Berlekamp) "Lower Bounds to Error Probability for Coding on Discrete Memoryless Channels I," *Information and Control*, Vol. 10 (1967), pp. 65-103. (Received Jan. 18, 1966.) Reprinted in D. Slepian, editor, *Key Papers in the Development of Information Theory*, IEEE Press, NY, 1974. Included in Part A.
- [123] (With R. G. Gallager and E. R. Berlekamp) "Lower Bounds to Error Probability for Coding on Discrete Memoryless Channels II," *Information and Control*, Vol. 10 (1967), pp. 522-552. (Received Jan. 18, 1966.) Reprinted in D. Slepian, editor, *Key Papers in the Development of Information Theory*, IEEE Press, NY, 1974. Included in Part A.
- [124] "The Fourth-Dimensional Twist, or a Modest Proposal in Aid of the American Driver in England," typescript, All Souls College, Oxford, Trinity term, 1978, 7 pp. + 8 figs. Included in this volume.
- [125] "Claude Shannon's No-Drop Juggling Diorama," *Juggler's World*, Vol. 34 (March, 1982), pp. 20-22. Included in Part B.
- [126] "Scientific Aspects of Juggling," Typescript, circa 1980. Included in Part B.
- [127] "A Rubric on Rubik Cubics," Typescript, circa 1982, 6 pp. Included in this volume.

[5]

43

COVER SHEET FOR TECHNICAL MEMORANDA
RESEARCH DEPARTMENTSUBJECT: The Use of the Lakatos-Hickman Relay in a
Subscriber Sender - Case 20878

ROUTING:

- 1 - Patent Dept. (letter 9/27/40)
2 - ~~B.W.Kendall~~, Case File
3 - T.C.Fry
4 - A.B.Clark
5 - B.D.Holbrook
6 - G.R.Stibitz
7 - G.V.King
8 - Miss Hanle

MM- 40-130-179
DATE August 13, 1940
AUTHOR C.E.Shannon
INDEX NO. S4.2

ABSTRACT

A study is made of the possibilities of using the Lakatos-Hickman type relay for the counting, registering, steering, and pulse apportioning operations in a subscriber sender. Circuits are shown for the more important parts of the circuit where it appears that the new type relay would effect an economy.

uuu

The Use of the Lakatos-Hickman Relay in a Subscriber Sender -
Case 20878

MM-40-130-179

August 13, 1940

MEMORANDUM FOR FILE

The Lakatos-Hickman type relay¹ using the relay springs as part of the magnetic circuit can be used as a very economical type of pulse counter and registration device. In fact, one such relay with twenty moving springs can count and register up to ten pulses, while the same operation requires at least five ordinary relays, and some standard circuits use as many as twenty to reduce the spring loading on the relays and the contact loading in the pulsing circuit. It has been suggested that this new type of relay might be used for some or all of the many counting, steering, and registration circuits in a subscriber type sender. The present memorandum gives some circuits for accomplishing this. The chief problem in the design of these circuits is that of performing the various translating operations necessary in converting the incoming pulses into group and brush selections, or P.C.I. pulses as the case may be, without using more contact elements than are available on the counting relay. Two different solutions are given here. The first was made as economical as possible but at the cost of one disadvantage. Under certain conditions of contact failure in the thousands or hundreds register the sender will connect the subscriber to an incorrect number rather than connecting to a tell-tale and giving him a busy signal. The second circuit, which we will call the positive action circuit², is designed to overcome this difficulty but does so at the expense of more contacts and wiring. Some compromise between these circuits may be the most desirable. The circuits by no means represent a complete sender. It appears that the problems connected with the office code (i.e. the first two or three digits) can be handled without much difficulty. At any rate these circuits will depend on the type of decoder used, and would represent a second stage in the design. We have therefore designed what might be called a "four digit sender" considering only the problems arising in the thousands, hundreds, tens and units digits. We also have omitted consideration of the parts of the circuit used for control and supervisory purposes, since these can be easily handled by existing circuits, and do not directly involve the new type relay. Our chief purpose is to

¹See "Circuit Analysis for Lakatos-Hickman Type Relay",
G. R. Stibitz, MM40-130-126, Jan. 15, 1940, Case 20878.

²This circuit was suggested by Mr. G. V. King

show that the new type counter contains sufficient contact elements for most of the steering and counting circuits of the subscriber sender. It is always possible to add more contacts at any stage in the new type counter by the arrangement of springs in Fig. 1, but this would be undesirable from the standpoint of standardization. At any rate it was found that even in the positive action circuit, only two stages in one register needed more contacts than are already available, and two additional ordinary relays were introduced here to carry the contact load.

It should be pointed out that an extremely simple and economical sender (i.e., much simpler than those given here) could be designed using the new type counter were it not for the peculiar translation codes involved. Thus if we could start "from scratch" and design translation codes particularly adapted to the characteristics of the new relay, the circuits could be made very simple indeed. Even using the existing codes which were constructed to simplify the present type circuits, the use of the new counter allows a remarkable simplicity and economy.

The circuits were designed by a combination of common sense and Boolean algebra methods. We will omit the details involved in their design. Although it is possible that a few superfluous elements remain, it is doubtful if they can be simplified very much.

Figure 2 is a block diagram of the proposed sender. In the present panel and crossbar senders, pulse counting is done in the same circuit for each digit and the numbers transferred from this counting circuit to a set of registering circuits, one for each digit, through an incoming steering chain. The registering circuits in the panel type sender consist of a set of five ordinary relays per digit, while in the crossbar system the A digit is registered on one or two verticals of a crossbar switch. In Figure 2, on the other hand, each digit has one of the new type counter relays which acts both as a pulse counter and as a register. The incoming steering chain steers the incoming pulses to the correct counter-register rather than steering the number recorded by the input pulse counter to a digit register. The input steering chain may or may not be one of the new type counters. The steering operation can be done with the new type counter, but it appears to require special devices, as for example polarized springs, in order to energize both windings of the register relays after receiving a digit. Even using the present type of steering chain a great simplification is possible, for only one wire, the pulsing lead, needs to be steered to the various digit registers, rather than the five leads of the present type sender. Another possibility is using a new type counter to count the groups of pulses and operate a set of relays S_A , S_B .

So, StH, SH, St, Sy which come in after the A, B, C, TH, H, T, and U digits are received and energize both coils of the corresponding registers.

After the digits are registered on the new type counters, these numbers are translated by means of the contact interconnections into the code corresponding to the incoming brush, incoming group, final brush, tens, and units selections, which are represented by a ground on one of the leads in the groups marked IB, IG, FB, T, and U, respectively. These groups of leads are connected in sequence to the revertive pulse counter by means of the revertive group counter. The revertive pulse counter will be one of the new type relays and is connected in such a way as to open the fundamental circuit and thus stop the revertive pulsing when it reaches the first ground. The revertive group counter or revertive steering chain, of course, steps ahead after each group of revertive pulses through the action of a slow release relay. This last steering operation cannot be done solely with one of the new type relays for it is necessary to steer ten leads in the tens and units digits. It could be done, however, with a new type counter in conjunction with four ordinary relays.

In the case of a call to a manual office the outputs of the digit registers are translated by a P.C.I. circuit into the correct P.C.I. codes. This circuit, too, can make use of the new type counter in the quadranting operation, i.e. in apportioning four quadrants to each of the four digits to be transmitted. This would be done with a sixteen stage counter (or if it is desirable to have all counters with ten stages, two of these could be connected "in series") replacing the present sequence switch.

Of course there must be an interlock between the incoming and revertive steering chains to prevent any selection being made before sufficient information has been received. This can be done by fairly standard methods.

A rough comparison can be made between the relay requirements of the present panel type sender and the design proposed here. Omitting parts of the circuit which would be substantially the same the requirements are listed below:

Operation	Present Panel Sender	Proposed Sender	
	Ordinary Relays	New Type Counters	Ordinary Relays
Input Counting	12	-	-
Input Steering	16	1	-
Registration	38	8	2
Revertive Counting	20	1	-
Revertive Steering	<u>10</u>	<u>1</u>	<u>5</u>
Total	96	11	7

In addition, a sequence switch is replaced by a new type counter. These figures are based on the positive action circuit. The other circuit uses 6 ordinary relays. This comparison of the numbers of relays involved shows only a small part of the saving, however. The wiring and fundamental method of operation of the new circuit is much simpler which tends both toward economy and, providing the new relay can be made sufficiently reliable, elimination of faults and errors.

It is a little more difficult to give a quantitative comparison of the proposed sender with the present crossbar type sender due to the differences in the types of circuit elements involved, but it appears that the saving would be of the same order of magnitude.

The new type counter with ten stages acts like a series of twenty relays which come in sequentially as the two coils of the relay are alternately energized. Thus after n pulses the first $2n$ relays are operated. If, after a series of pulses only one of the two coils on a counter remains energized we can only be sure of the contacts on that side. It was found that under these conditions the number of contacts available was far too small in all of the four registers for the various translating operations necessary. We have therefore assumed the steering circuit should be designed in such a way as to energize both coils of a counter after it has received its series of pulses.* This insures the contacts on both sides and each stage then has the equivalent of two transfer contacts and two additional contacts somewhat similar to a switchhook connection. Thus each stage may be considered as a relay with the contacts available indicated in Figure 3. Our circuit diagrams are drawn from this point of view.

For the convenience of the reader we will list the various translation codes used in the sender. The incoming brush selection depends only on the thousands digit and is given by the following table:

Thousands Digit	Incoming Brush Selection
0, 1	0
2, 3	1
4, 5	2
6, 7	3
8, 9	4

*See the memorandum "Circuit Arrangement for Counting Relay with Mechanically Independent Contact Springs", by B. D. Halbrook, MM-40-130-149, July 5, 1940, Case 22108-1.

The incoming group selection depends on both the hundreds and thousands digits and is given by the following:

Thousands Digit	Hundreds Digit	Incoming Group Selection
even	< 5	0
even	≥ 5	1
odd	< 5	2
odd	≥ 5	3

The final brush selection depends only on the hundreds digit. We have the following code:

Hundreds Digit	Final Brush Selection
0, 5	0
1, 6	1
2, 7	2
3, 8	3
4, 9	4

It should be remembered that an incoming brush, incoming group, or final brush selection of n corresponds to $n + 1$ reverberative pulses. The same remark applies to the tens and hundreds selection.

Digits are sent to a call indicator by series of positive and negative pulses, four for each digit. Two different codes are used for this, one for the thousands digit and the other for the hundreds, tens, and units. The thousands code is an additive one based on the numbers 1, 2, 4, and 8 as follows:

P.C.I. Code for Thousands Digit

Thousands Digit	Quadrant			
	I	II	III	IV
1	0	0	0	-
2	-	0	0	0
3	-	0	0	-
4	0	-	0	0
5	0	-	0	-
6	-	-	0	0
7	-	-	0	-
8	0	0	-	0
9	0	0	-	-
0	0	0	0	0
Corresponding Additive Numbers	2	4	8	1

The sum of the numbers corresponding to the columns in which a digit has the symbol - gives that digit, hence the additive property of the code. In this table I, II, III, and IV refer to the four pulses or quadrants. In the first and third quadrants 0 represents a ground and a - represents a positive pulse. In the even quadrants 0 means a light negative pulse and the -, a heavy negative pulse. We have chosen this representation of the code for comparison with the P.C.I. circuit in which four leads are grounded or not in accordance with the above table. Thus if the digit 3 is registered in the thousands place, leads II and III in a group I, II, III, IV are grounded. The presence or absence of these grounds are translated into positive or negative pulses by two relays TS and RS.

The hundreds, tens, and units P.C.I. code is also additive based on the numbers 1, 2, 4, 5. Using the same conventions it is represented by the following table:

P.C.I. Code for Hundreds, Tens, and Units Digits

H, T, or U Digit	Quadrant			
	I	II	III	IV
1	-	0	0	0
2	0	-	0	0
3	-	-	0	0
4	0	0	-	0
5	0	0	0	-
6	-	0	0	-
7	0	-	0	-
8	-	-	0	-
9	0	0	-	-
0	0	0	0	0
Corresponding Numbers	(1)	(2)	(4)	(5)

The circuit for the tens or units register is shown in Figure 4. The operation is quite obvious. In the case of a full mechanical call, if 6 for example were dialed in the tens place, the first six relays are locked in, which places a ground on the lead marked 6. These are connected through the revertive steering chain to the revertive counter which reaches this ground after the seventh revertive pulse. The presence of this ground operates a relay which opens the fundamental circuit and stops the pulsing. A ground is also put on leads II and III for a P.C.I. call. The operation of the P.C.I. circuit will be described later. The thousands and hundreds register is shown in Figure 5 for the positive action circuit and in Figure 6 for the more economical circuit. In Figure 5, many of the contacts do double duty, translating both for P.C.I. and full mechanical calls. This is done through a relay P which is operated for a manual call and not for a mechanical call. In the hundreds register there were not enough contacts available in the fifth and tenth stages.

The relays R and S are used to carry part of the contact load. This circuit is designed so that one and only one of the IB, IG, and FB leads is grounded for a given number. In case of a contact failure none would be grounded and the corresponding commutator would supposedly go to a telltale. In the circuit of Figure 6, on the other hand, more than one of the IB, IG, or FB leads may be grounded at the same time. Thus if the thousands digit is 8, both 8 and 4 in the IB group are grounded. If the back contact on 8 failed the revertive pulse counter would not stop the pulsing action at brush 3 as it should but would go on to the fourth brush. However, this circuit is considerably simpler than Figure 5, and does not appear worse from the standpoint of possible wrong numbers than the present type of sender.

The P.C.I. circuit is shown in Figure 7. X is a relay which is operated in the odd quadrants and not in the even quadrants. TS and RS are relays whose windings are connected sequentially through the P.C.I. impulse chain to first the thousands P.C.I. leads I, II, III, and IV, then the hundreds, etc. according to the following table:

Pulsing Stage			TS	RS
Th Digit	(1	X	Th I	Th II
	(2	-	Th III	Th II
	(3	X	Th III	Th IV
	(4	-	H I	Th IV
H Digit	(5	X	H I	H II
	(6	-	H III	H II
	(7	X	H III	H IV
	(8	-	T I	H IV
T Digit	(9	X	T I	T II
	(10	-	T III	T II
	(11	X	T III	T IV
	(12	-	U I	T IV
U Digit	(13	X	U I	U II
	(14	-	U III	U II
	(15	X	U III	U IV
	(16	-	-	U IV

In the odd quadrants X is operated, placing a ground on the fundamental ring (FR). The fundamental tip (FT) is connected through X to either ground or positive battery according as TS is operated or not. This depends of course on the condi-

tion of the P.C.I. lead to which TS is connected at the time. Similarly in the even quadrants light or heavy voltage is applied to FR according to the condition of RS while FT is grounded.

Figure 8 shows the revertive steering chain and revertive pulse counter.

C. E. SHANNON

C. E. S.

Pulsing Stage		TS		RS	
1	{	X	TS I	TS I	TS I
			TS III	TS II	TS II
			TS V	TS IV	TS IV
			TS VII	TS VI	TS VI
2	{	X	TS I	TS I	TS I
			TS III	TS II	TS II
			TS V	TS IV	TS IV
			TS VII	TS VI	TS VI
3	{	X	TS I	TS I	TS I
			TS III	TS II	TS II
			TS V	TS IV	TS IV
			TS VII	TS VI	TS VI
4	{	X	TS I	TS I	TS I
			TS III	TS II	TS II
			TS V	TS IV	TS IV
			TS VII	TS VI	TS VI
5	{	X	TS I	TS I	TS I
			TS III	TS II	TS II
			TS V	TS IV	TS IV
			TS VII	TS VI	TS VI
6	{	X	TS I	TS I	TS I
			TS III	TS II	TS II
			TS V	TS IV	TS IV
			TS VII	TS VI	TS VI
7	{	X	TS I	TS I	TS I
			TS III	TS II	TS II
			TS V	TS IV	TS IV
			TS VII	TS VI	TS VI
8	{	X	TS I	TS I	TS I
			TS III	TS II	TS II
			TS V	TS IV	TS IV
			TS VII	TS VI	TS VI

In the odd quadrants X is operated, placing a ground on the fundamental ring (X). The fundamental slip (Y) is connected through X to either ground or positive battery according as X is operated or not. This depends of course on the condi-

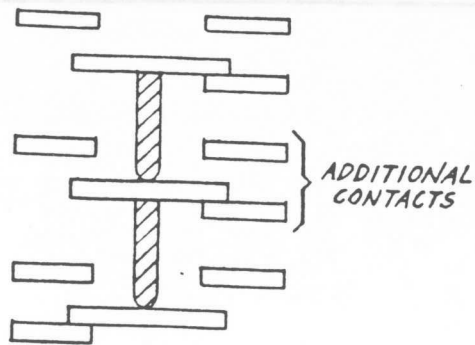


FIG. 1

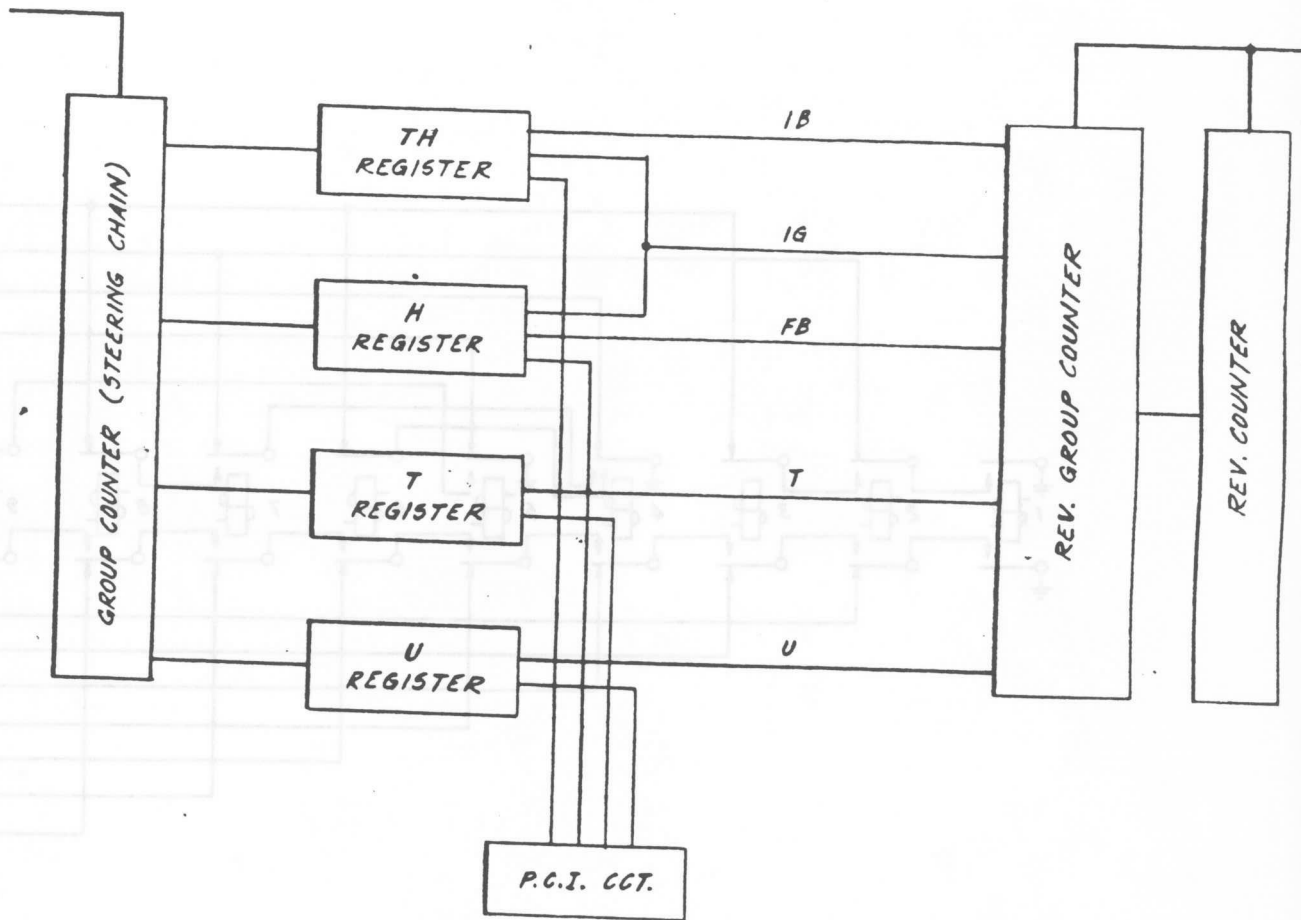


FIG. 2

BLOCK DIAGRAM OF SENDER

ISSUE 1 8-22-40

APPL.	DR.	CH.	ENG.	TITLE
	B.C.S.		C.E.S.	
	TR.			

SCALE
BELL TELEPHONE LABORATORIES, INC., NEW YORK

ES

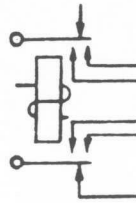


FIG. 3

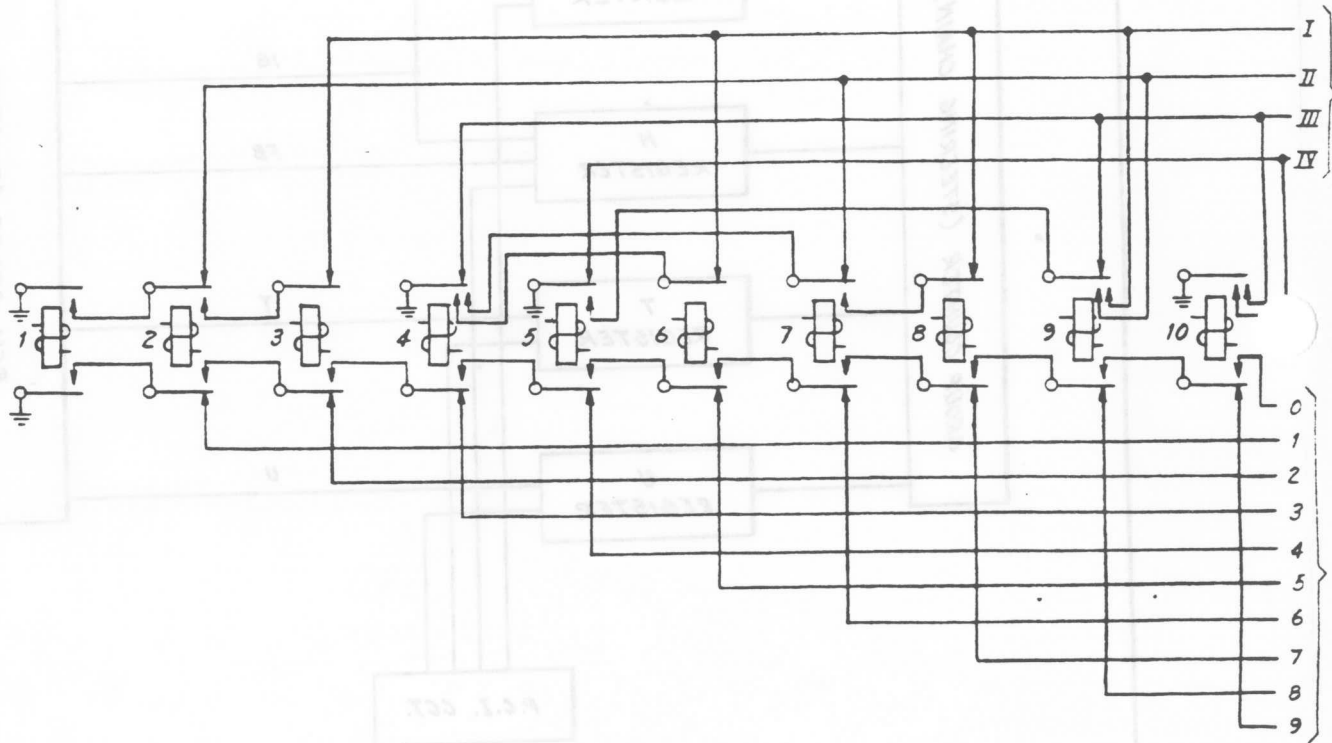


FIG. 4

TENS OR UNITS REGISTER

CH.		TITLE
DR.	ENG.	
B.C.S.	C.E.S.	
TR.		
APPL.		SCALE
		BELL TELEPHONE LABORATORIES, INC., DET
		ES

Issue 1 8-23-40

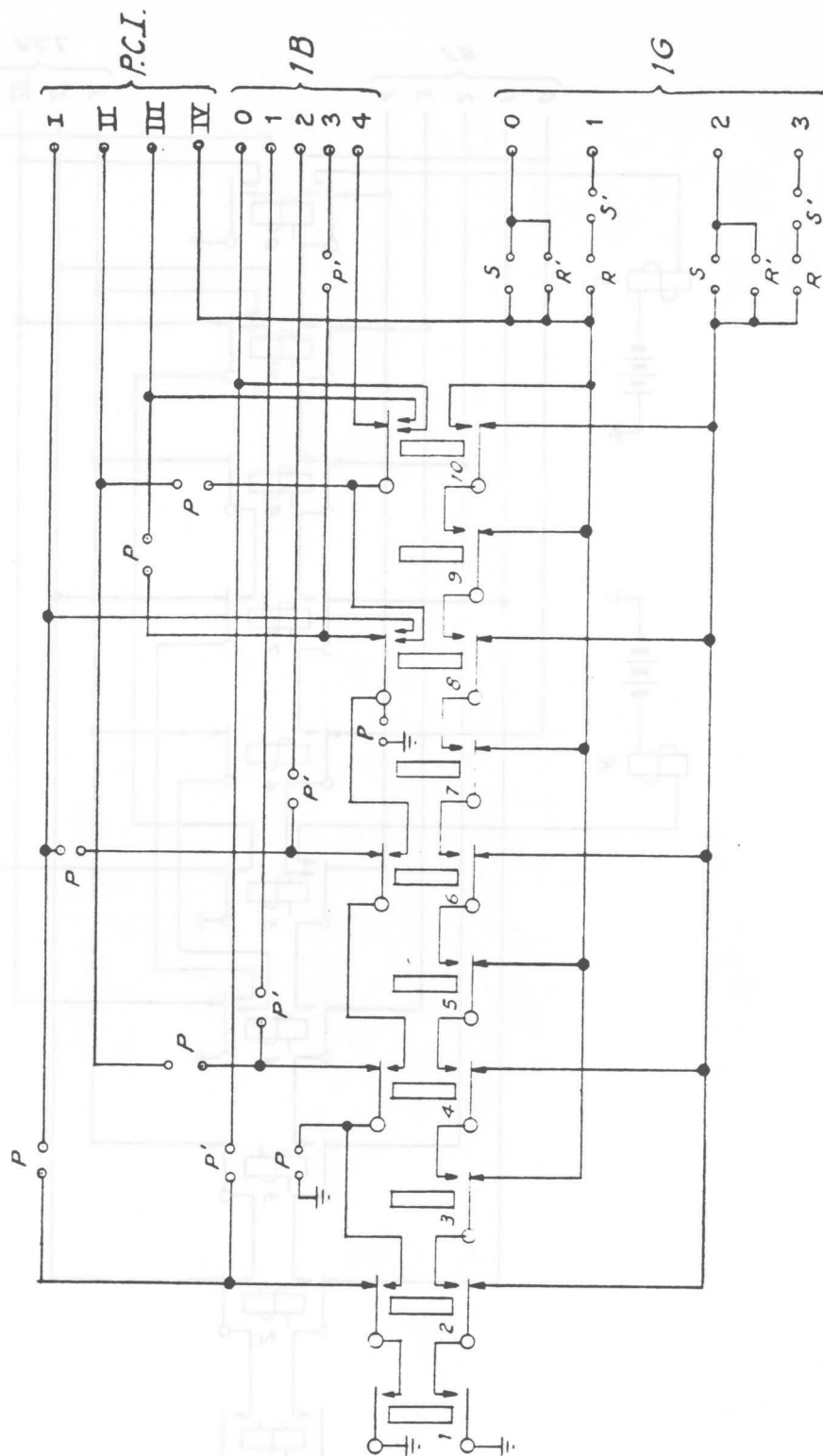


FIG. 5a
THOUSANDS REGISTER
POSITIVE ACTION CIRCUIT

APPL.	DR.	CH.	TITLE
	F. M. T.	ENG.	
	TR.	G. E. S.	
SCALE			ES
BELL TELEPHONE LABORATORIES, INC., NEW YORK			

10-11-40
 10-11-40
 10-11-40

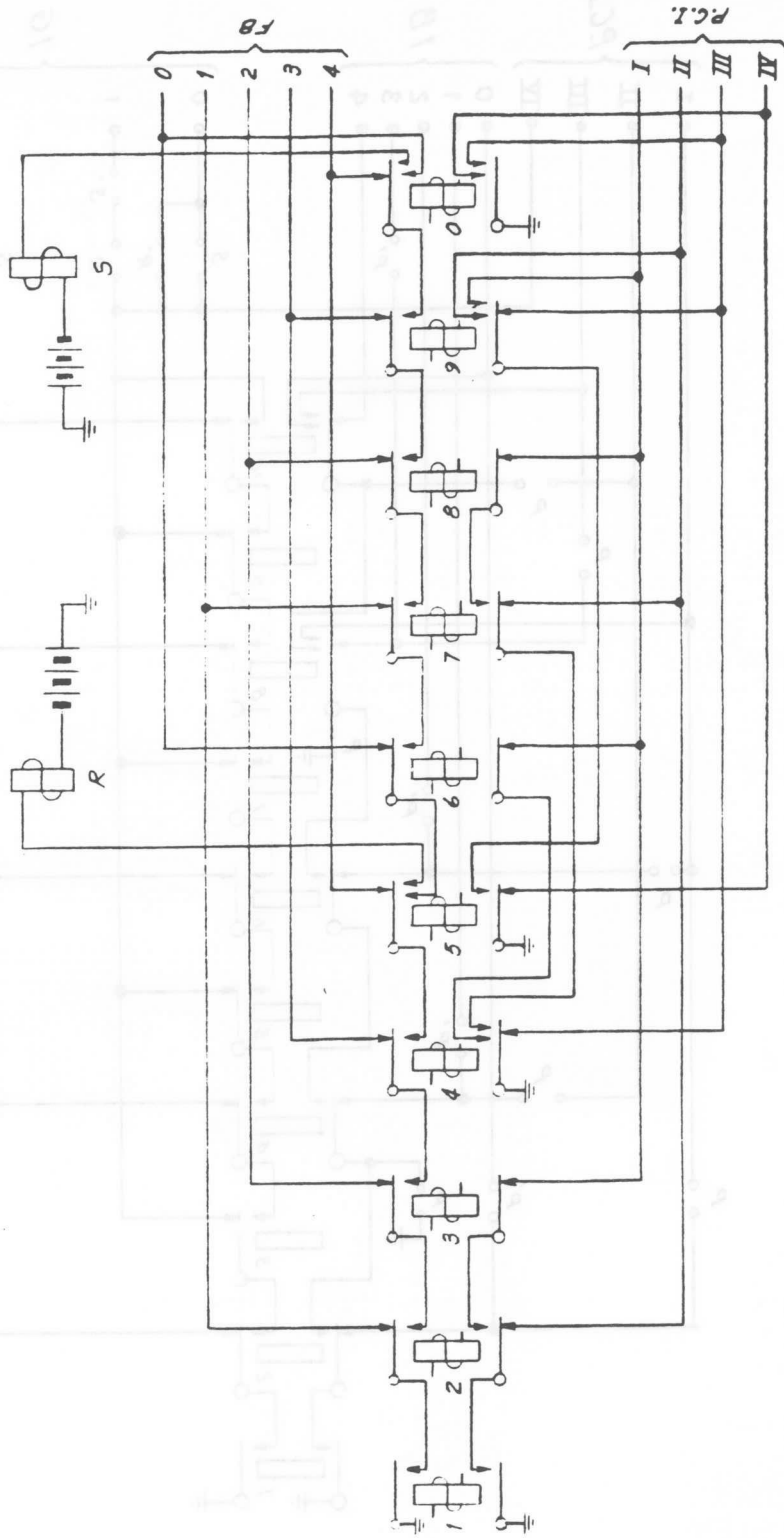


FIG. 5b
 HUNDREDS REGISTER, POSITIVE ACTION CIRCUIT

ISSUE 1 8-22-40

APPL.	DR.	CH.
	B.C.S.	ENG.
	TR.	C.E.S.

TITLE	
SCALE	
BELL TELEPHONE LABORATORIES, INC., NE	
ES	

ISSUE 1 8-23-40

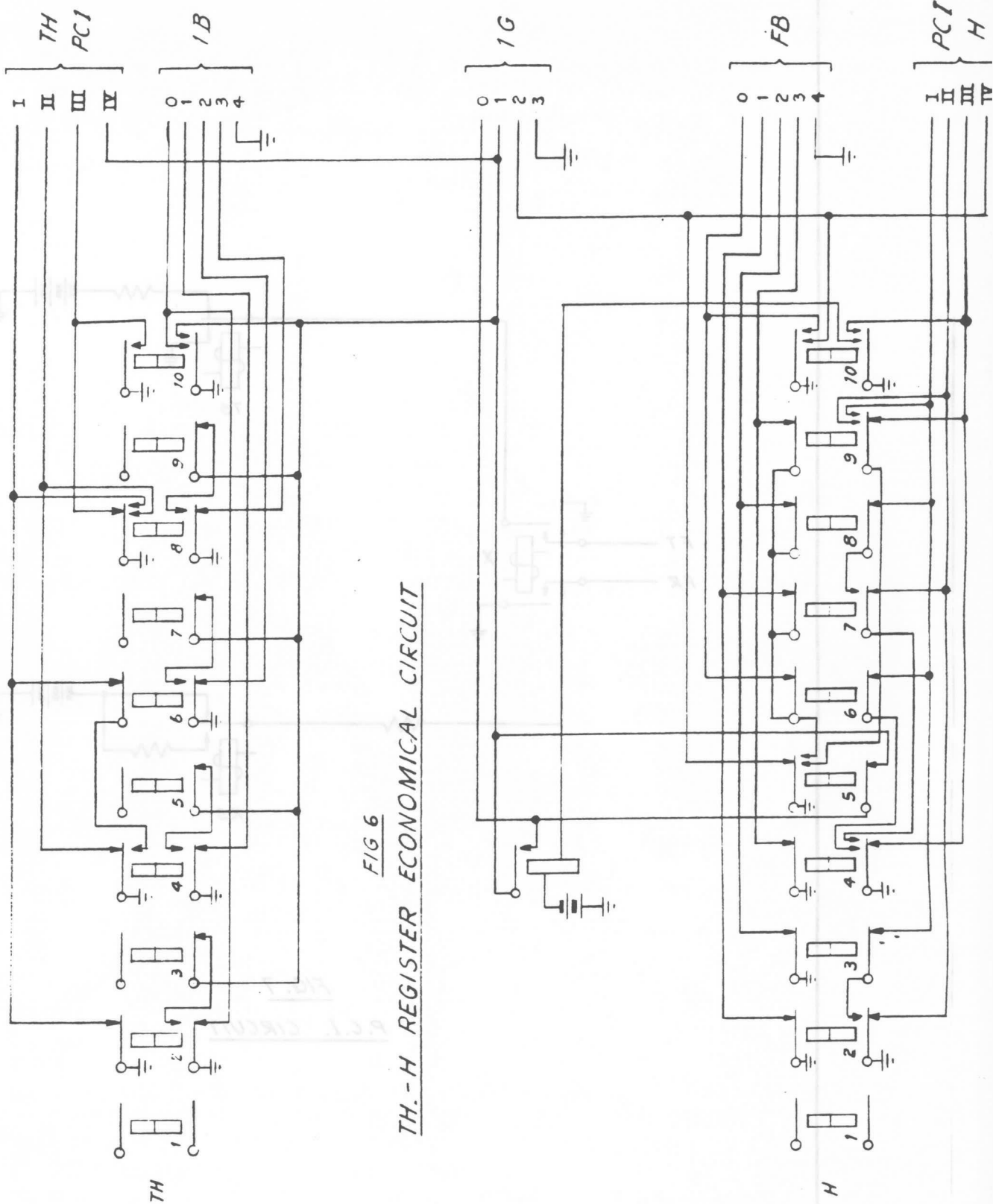


FIG 6
TH.-H REGISTER ECONOMICAL CIRCUIT

APPL.	DR.	CH.	TITLE
	F.M.T.	ENG.	
	TR.	C.E.S.	
SCALE			ES
BELL TELEPHONE LABORATORIES, INC., NEW			
PRINTED IN U.S.A.			

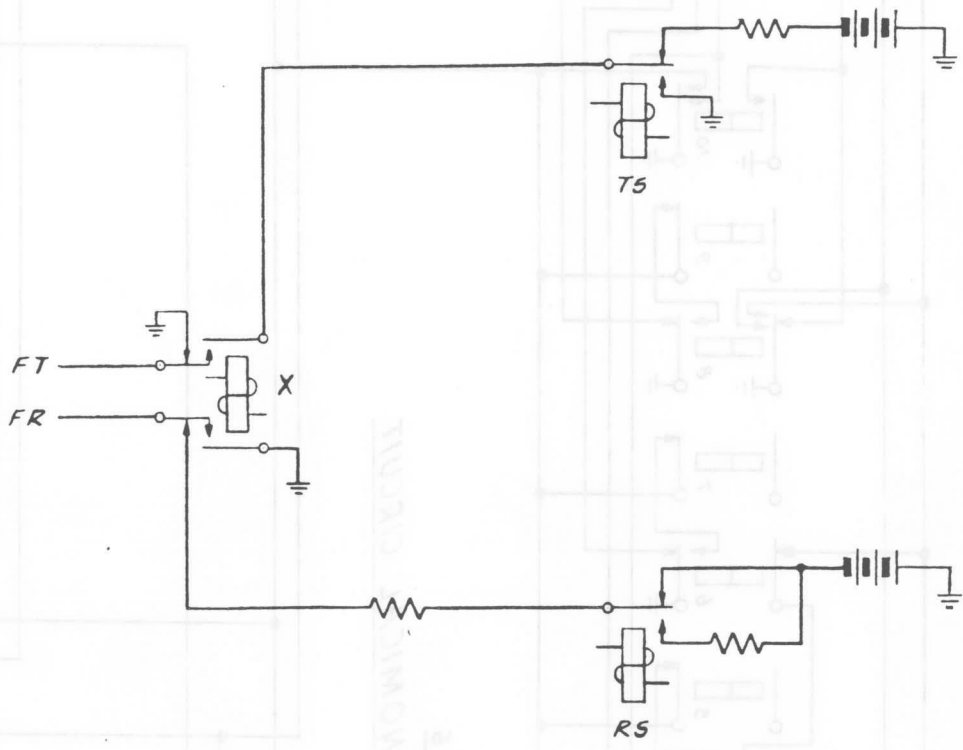


FIG. 7
P.C.I. CIRCUIT

Issue / 8-23-40

APPL.		DR.	CH.	TITLE	
		B.C.S.		ENG.	
		TR.		C.E.S.	
				SCALE	
				BELL TELEPHONE LABORATORIES, INC., NEW	
				ES	

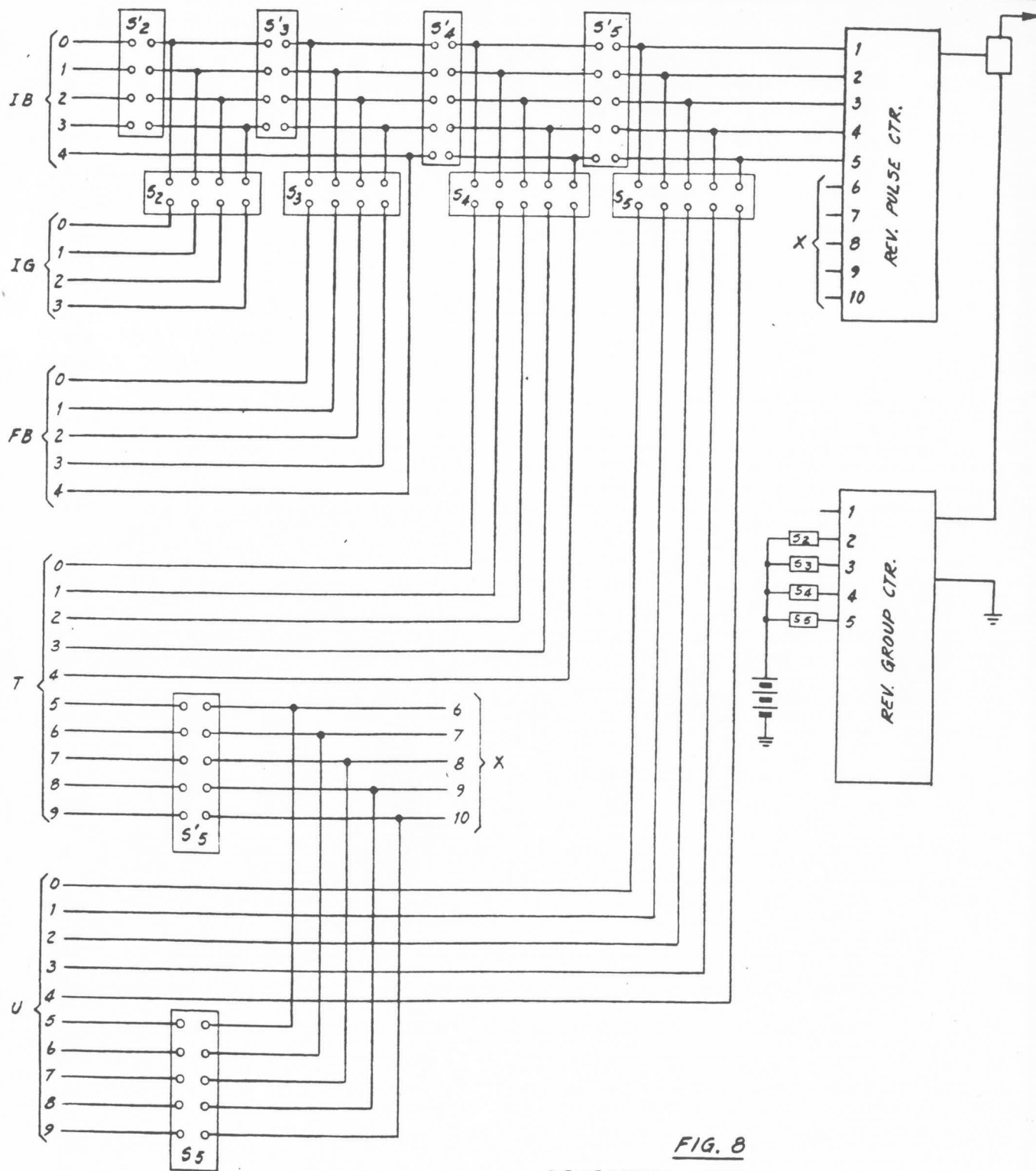


FIG. 8
REVERTIVE STEERING CHAIN

ISSUE 1 8-23-40

APPL.	DR.	CH.	TITLE
	D.C.S.	ENG.	
	TR.	C.C.S.	
SCALE			BELL TELEPHONE LABORATORIES, INC., NEW YORK
ES			

[7]

as True Report

7/14-46

[Handwritten signature]

...the document for the
...the morning of the
...M.E.C. 50:31 32). The transcrip-
...the document for the revelation
...to any person or an individ-
...red person is prohibited by law.

ND C ~~ac~~ / 05

A STUDY OF THE DEFLECTION MECHANISM
AND SOME RESULTS ON RATE FINDERS

by

Claude E. Shannon

THIS IS A FINAL REPORT
UNDER CONTRACT

100-100000-100000

THE CO-

CONFIDENTIAL

SUMMARY OF THE MOST IMPORTANT RESULTS

1. The deflection mechanism may be divided into three parts.

The first is driven by two shafts and has one shaft as output, which feeds the second part. This unit has a single shaft output which serves as input to the third part, whose output is also a single shaft, used as the desired azimuth correction.

2. The first unit is a simple integrator. Its output rate is

$$\dot{y} = \sum a \frac{R_0}{R_p} t_p$$

3. The second part is the same circuit as previous rate finders.

Its presence appears to be detrimental to the operation of the system from several standpoints. The output e of this part satisfies:

$$e = x + y$$

$$\frac{R_1}{L_1} x + \dot{x} = y$$

4. The third and most important part of the machine satisfies

$$S q + R \dot{q} + \frac{L}{\sqrt{1 - q^2}} \ddot{q} = e$$

in which:

e = an input forcing function which except for transients in the second part and other small effects is the function whose rate is to be found.

\dot{q} = the rate of e as found by the device. The output of the mechanism is $\sin^{-1} \dot{q}$.

R, L, S are positive constants depending on the gear ratios, etc. in the machine.

5. The mechanism therefore acts like an R, L, C circuit in which the differential inductance is a function of the current,

$$\frac{L}{\sqrt{1 - \dot{q}^2}}.$$

The system can be critically damped for differential displacements near at most two values of the current.

6. Omitting the effect of backlash, the system is stable for any initial conditions whatever, with a linear forcing function, $e = At + B$. It will approach asymptotically and possibly with oscillation a position where \dot{q} is proportional to \dot{e} . An error function can be found which decreases at a rate $-R(\dot{q} - \dot{q}_0)^2$ \dot{q}_0 being the asymptotic value of \dot{q} .

7. If the system is less than critically damped ordinary gear play type of backlash can and will cause oscillation. This includes play in gears, adders, lead screws, rack and pinions and looseness of balls in the integrator carriages. The oscillation is not unstable in the sense of being erratic, or growing

without limit, but is of a perfectly definite frequency and amplitude. This type of backlash acts exactly like a peculiar shaped periodic forcing function. Approximate formulas for the frequency and amplitude of the oscillation are

$$f_0 = \frac{1}{2\pi} \sqrt{\frac{1}{LD^2} + \frac{R^2}{4L^2}}$$

and

$$\frac{2}{\pi} B_1 + 4 f_0 B_2$$

$$I = \frac{\frac{2}{\pi} B_1 + 4 f_0 B_2}{\sqrt{R^2 + \left(\omega_0 LD - \frac{1}{\omega_0 C}\right)^2}}$$

B_1 and B_2 being the amounts of backlash in the two driven shafts as measured in a certain manner.

8. Elastic deformations of shafts and plates can be divided into two parts. One is exactly equivalent to the gear type of backlash and may be grouped with B_1 and B_2 above. The other has the effect of altering the parameters R , L , S of the circuit and also adding higher order derivatives with small coefficients. This will slightly alter the time constant and the natural frequency of the system.

9. The manner in which the arcsin function is obtained seems to me distinctly disadvantageous to the operation of the system for a number of reasons, chiefly since to eliminate backlash

oscillation it requires high overdamping near $\dot{q} = 0$ and this slows down the response for low target speeds.

10. The general problem of rate finding and smoothing is considered briefly from two angles - as a problem in approximating a certain given transfer admittance and as a problem in finding the form of a differential equation. The first method based on a linear differential equation leads to tentative designs which I think would be an improvement over the present one. The second method indicates the possibility of still more improvement if non-linear equations can be satisfactorily analyzed.

ANALYSIS OF THE DEFLECTION MECHANISM

General Considerations. The deflection mechanism is a device designed to find δ_1 mechanically from the formula

$$\sin \delta_1 = \sum_a \frac{R_0}{R_p} t_p$$

having one shaft whose rate of turning is \sum_a and another whose angular position is $\frac{R_0}{R_p} t_p$, giving δ_1 as the position of a shaft. The system is also supposed to smooth out small errors in \sum_a .

The mechanism, as actually constructed, is shown in Figure 1. By a rearrangement of adders, it may be drawn as shown in Figure 2. Incidentally, the device of rearranging and combining

adder units is frequently useful in studying these systems. In this case it both clarifies the physical operation and simplifies the mathematical analysis. The box IV on the right of Fig. 1 represents two adders with, essentially, a common shaft. The output is equal to the sum of the inputs with the indicated signs prefixed. A variable associated with a shaft represents the angular position of that shaft unless specifically stated otherwise. Gears are omitted from the diagram but included as coefficients in the equations. It may also be worthwhile to point out that the best method of setting down the equation of such a system is usually the following:

1. Considering only the integrators and function devices, label the various shafts using the minimum number of variables, working backward from driven to driving shafts. Thus if the output of an integrator is labeled z , its displacement is \dot{z} (assuming constant disk rate). If the output of an x to $\ln x$ gear is $\sin u$, its input is $e^{\sin u}$. Working backwards gives the differential instead of the integral form of the equation.

2. Now concentrate on the adders, grouping together as many as possible, and write the equations of constraint. These will be the equations of the system.

I find the use of electrical analogues very useful in understanding these devices and have used throughout a notation which emphasizes this idea.

As the machine is drawn in Fig. 2, it consists of three independently operating units. The output of the first is a single shaft serving as input to the second, the output of the second a single shaft feeding the third, and the output of this being a shaft used as δ_1 .

The operation is roughly as follows: Integrator I multiplies its disk rate by its displacement, so that the rate of turning of its output is $\dot{y} = \frac{R_o}{R_p} t_p \sum a$. The actual position of this y shaft can carry no significance. It is

$$y = \int_0^t \frac{R_o}{R_p} t_p \sum a \, dt + y_0$$

a variable which depends on the entire previous history of the sighting telescopes to say nothing of possible integrator slippage. At two different times, with a target at the same position and speed, this shaft would have entirely different angular positions but the same rate of turning.

The output of integrator I feeds into the middle part of the system which is exactly the rate finder, of most older directors. This part of the device seems to me not only superfluous but actually detrimental to the operation. It is equivalent to an R, L, circuit (Fig. 3) with impressed voltage y and output x, the voltage across the inductance

3. A small response $h(t)$ for the function $g(t)$.

High frequencies in $g(t)$ appear practically undiminished and in the same phase in $h(t)$ since the impedance is high compared to R .

Thus

$$x = \frac{L_1}{R_1} A + K e^{-\frac{R_1}{L_1} t} + h(t)$$

In adder III, x is added to y in equal proportions to give e .

$$e = y + \frac{L_1}{R_1} A + K e^{-\frac{R_1}{L_1} t} + h(t)$$

As we pointed out above, y already contains an irrelevant additive constant, so the addition of another, $\frac{L_1}{R_1} A$ which happens to be proportional to the target rate is of no possible significance. The term $K e^{-\frac{R_1}{L_1} t}$ certainly is only detrimental being an unwanted transient. For a time I thought that the reason for the middle part of the machine was the final term $h(t)$. For high frequencies this is approximately $g(t)$, and might be used to buck out these high frequency following errors, much as was done in some early radio circuits to reduce a-c hum. However, a study of the design diagrams shows that the two error functions are actually in phase as I have indicated in the equation, so that these high frequency errors are added, making the situation worse. Even if the phase of x were reversed on entering adder III, I think it

doubtful whether the presence of this part of the system would be justifiable. It would be necessary to show that the frequencies were high enough so that the two actually did cancel, and also that the disadvantages of the transient term did not overcome the advantages obtained. Note that the middle part can function in no way as a rate finder. The right hand part of the machine does its own rate finding as we will see, and the rate found by the middle part could not possibly be used because of the undetermined constant in y .

We proceed now to the third part of the machine which is the major concern of the study. Concentrating on the adder IV, the equation of the system is obviously

$$L \frac{d}{dt} \sin^{-1} \dot{q} = e - S q - R \dot{q}$$

or

$$S q + R \dot{q} + \frac{L}{\sqrt{1 - \dot{q}^2}} \ddot{q} = e$$

This is the equation of a series R, L, C , circuit with the inductance a function of the current passing through it. Inductance may be defined by the Lagrangian equations or by

$$e_L = \frac{d}{dt} \Lambda i = \frac{d}{dt} \Lambda \dot{q}$$

and it is clear from the above equation that

$$\Lambda i = L \sin^{-1} i$$

$$\text{or } \Lambda = L \frac{\sin^{-1} i}{i}$$

This function varies as shown in figure 4. For our work, however a more useful parameter is what is sometimes called the differential inductance L_D which may be defined by

$$e_L = L_D \frac{di}{dt}$$

so that in our case

$$L_D = \frac{L}{\sqrt{1 - q^2}}$$

This inductance is useful when we have an equilibrium current q_E and are considering the effect of small variations about this equilibrium. Omitting second order terms the system will be equivalent to one with constant R , L , C parameters, the inductance being taken as L_D . The variation of L_D with current is shown in figure 5.

- 01 -

$$\dot{\theta} \wedge \frac{b}{25} = \dot{\theta} \wedge \frac{b}{35} = \dot{\theta}^2$$

and it is clear from the above equation that

$$\dot{\theta} \wedge \frac{b}{25} = \dot{\theta} \wedge \frac{b}{35} = \dot{\theta}^2$$

This function varies as shown in figure 4. For our work, however,

The action is the opposite of that of a "swinging" choke where, because of saturation, the differential inductance decreases with large currents.

The mechanical idea behind the operation of this system is quite simple. Suppose shaft e to be turning at a constant rate. The system will be in equilibrium if the displacement of integrator V is such as to make its output feeding into the adder equal and opposite to e, and the displacement of integrator VI at zero. Under these conditions, shaft q measures the rate of e and shaft V, the output of the device, the arcsin of this rate. If the rates are not correct, the adder changes the second derivative shaft in such a direction as to equalize the rates. The q shaft serves as a damper to prevent continual oscillation about the equilibrium position.

MATHEMATICAL THEORY (Backlash not Present)

Differential Operation

If e is turning at a constant rate and the system is at equilibrium, and then a small differential disturbance is applied to the system, it will clearly respond very nearly like an R, L, C, circuit with constant parameters, the inductance used being the differential inductance for the equilibrium current

$$L_D = \frac{L}{\sqrt{1 - \dot{q}_E^2}}$$

Such a system has a time constant of

$$T = \frac{2 L_{eff}}{R} = \frac{2L}{R \sqrt{1 - \dot{q}_E^2}}$$

It is critically damped if

$$R^2 = 4 L_{eff} S = \frac{4L S}{\sqrt{1 - \dot{q}_E^2}}$$

which, of course, only occurs at

$$\dot{q}_E = \pm \sqrt{1 - \frac{16 L^2}{R^2 C^2}}$$

For values of \dot{q} greater in absolute value than this, the system is oscillatory, for values less, overdamped.

Proof of General Stability with Linear e

In proving the stability of this system, I have used a method which may be new in some respects. It was suggested by the fact that in a non-dissipative mechanical system, the potential energy U is a minimum at a point where the system is differentially stable, and the method is, in a sense, a generalization of that criterion. It is not, however, limited to differential stability, or to non-dissipative systems. Since the method may be of use in other investigations of this type, I will first describe it in general terms.

Suppose we have a differential equation system in which n variables and derivatives may be specified independently in the initial conditions. We will say that the system is stable for all initial conditions and all driving functions if any two solutions of the system with the same driving functions approach each other in the sense that

$$\lim_{t \rightarrow \infty} \sum_{i=1}^n |x_i - y_i| = 0$$

where $x_1(t), x_2(t) \dots x_n(t)$ is one solution and $y_1(t) \dots y_n(t)$ the other. If this limit is zero for certain types of driving functions, we will say the system is stable for these functions.

Theorem: If a continuous function $Q(x_1 \dots x_n, y_1 \dots y_n, t)$ can be found having the following properties

1. $Q \geq 0$ for all x_i, y_i, t , the equality holding if and only if $x_i = y_i$.

2. $\frac{dQ}{dt} \leq 0$ at all times, when the x_1 and y_1 are solutions of the system, with the same driving function.

3. It is impossible for Q to remain indefinitely $\geq A > 0$.

Then the system is completely stable.

For the function Q is non-increasing but always ≥ 0 and must therefore approach a limit $A \geq 0$ as $t \rightarrow \infty$, but by 3. $A > 0$ is impossible, hence $A = 0$, and each $|x_1 - y_1| \rightarrow 0$.

Conversely, it can be shown that if only a single forcing function is involved, and the system is stable for this function, a Q exists of the type described.

Roughly, the method is to find a "distance" or "error" function Q between two solutions which is zero only when the solutions are identical and which always decreases.

As an example of this method it is easy to prove the complete stability of the ordinary R, L, C, circuit with constant parameters without solving the equation. The differential equation is

$$Sq + R\dot{q} + Lq = e$$

and we choose q and \dot{q} as coordinates. Let two solutions be q_1 ,

\dot{q}_1 and q_2 , \dot{q}_2 and consider the function $Q = \frac{S}{2} (q_1 - q_2)^2 + \frac{L}{2} (\dot{q}_1 - \dot{q}_2)^2$.

Condition 1 is obviously satisfied. Now

$$\begin{aligned} \frac{dQ}{dt} &= S(q_1 - q_2)(\dot{q}_1 - \dot{q}_2) + L(\dot{q}_1 - \dot{q}_2)(\ddot{q}_1 - \ddot{q}_2) \\ &= -R(\dot{q}_1 - \dot{q}_2)^2 \leq 0 \end{aligned}$$

$$Q(q, \dot{q}, t) = \frac{S}{2} \left(q - \frac{At}{S} - \frac{B}{S} + \frac{RA}{S^2} \right)^2 - L \left(\sqrt{1 - \dot{q}^2} + \frac{A}{S} \sin^{-1} \dot{q} \right) + L \left(\sqrt{1 - \left(\frac{A}{S} \right)^2} + \frac{A}{S} \sin^{-1} \frac{A}{S} \right)$$

obviously the minimum of Q with respect to q occurs at

$$q = \frac{At}{S} + \frac{B}{S} - \frac{RA}{S^2}$$

Also

$$\frac{\partial Q}{\partial \dot{q}} = L \frac{\dot{q} - \frac{A}{S}}{\sqrt{1 - \dot{q}^2}}$$

which vanishes only for $\dot{q} = \frac{A}{S}$. It is readily verified that this is a minimum, and that Q is zero at this point for any t . Now

$$\begin{aligned} \frac{dQ}{dt} &= \frac{\partial Q}{\partial q} \dot{q} + \frac{\partial Q}{\partial t} + \frac{\partial Q}{\partial \dot{q}} \ddot{q} \\ &= S \left(q - \frac{At}{S} - \frac{B}{S} + \frac{RA}{S^2} \right) \left(\dot{q} - \frac{A}{S} \right) + L \frac{\dot{q} - \frac{A}{S}}{\sqrt{1 - \dot{q}^2}} \ddot{q} \end{aligned}$$

and

$$\ddot{q} = \frac{\sqrt{1 - \dot{q}^2}}{L} (At + B - Sq - R\dot{q})$$

if q and \dot{q} satisfy

$$Sq + R\dot{q} + \frac{L}{\sqrt{1 - \dot{q}^2}} = At + B.$$

Hence

$$\begin{aligned}\frac{dQ}{dt} &= (Sq - At - B + \frac{RA}{S}) (\dot{q} - \frac{A}{S}) \\ &\quad - (\dot{q} - \frac{A}{S})(At + B - Sq - R\dot{q}) \\ &= (\dot{q} - \frac{A}{S})(\frac{RA}{S} - R\dot{q}) \\ &= -R(\dot{q} - \frac{A}{S})^2 \leq 0\end{aligned}$$

Note that this rate is identical with that found in the linear case. Incidentally, it was by working backward from this rate that a suitable function Q was first found.

For Q to approach a limit $K > 0$, it is necessary for \dot{q} to approach zero, and q therefore, to approach a linear function of t differing by a constant from its equilibrium value. But from the original differential equation \ddot{q} must approach a constant different from zero, which contradicts $\dot{q} \rightarrow 0$. This does not however, quite complete the stability proof due to a certain mechanical peculiarity of the system. Let us plot the equilevel lines of Q against axes $X = (q - \frac{At}{S} - \frac{B}{S} + \frac{RA}{S^2})$ and $Y = \dot{q}$. (Figure 6).

$$\left(\frac{d}{dt} - \dot{\phi}\right) \left(\frac{d}{dt} + \frac{2\dot{\phi}}{3}\right) (2\phi - 3x - 3 + p\phi) = \frac{4\dot{\phi}}{3}$$

$$\left(\frac{d}{dt} - p\phi - 3 - 2\dot{\phi}\right) \left(\frac{d}{dt} - \dot{\phi}\right) =$$

$$\left(\frac{d}{dt} - \frac{2\dot{\phi}}{3}\right) \left(\frac{d}{dt} - \dot{\phi}\right) =$$

$$0 \geq \frac{2}{3} \left(\frac{d}{dt} - \dot{\phi}\right) \dot{\phi} =$$

Note that this rate is identical with that found in the linear case.

Incidentally, it was by working backward from this rate that a

suitable function Q was first found.

For Q to approach a limit $K > 0$, it is necessary for \dot{Q}

to approach zero, and Q therefore, to approach a linear function

of t differing by a constant from the equilibrium value. But from

the original differential equation \dot{Q} must approach a constant different

from zero, which contradicts $\dot{Q} = 0$. This does not however, preclude

The x to $\sin x$ gear in the actual mechanism has a limited

movement, and is prevented from going too far by a slip clutch and

stop. If $|\dot{q}| = K$, the stop prevents $|\dot{q}|$ from increasing any more.

The original equation is replaced by

$$\dot{q} = \begin{cases} -K \\ K \end{cases} \quad q = \pm Kt + C$$

until the pressure on the stop reverses. So far we have shown that

under the original equation Q always decreases. In terms of our

plot this means that if we start a solution inside the curve marked C,

the solution will certainly converge to the equilibrium position, for

the solution can never "escape" from C and hit one of the two lines

$q = \pm K$, where the differential equation changes. When we are not on

one of these lines a solution will, in fact, spiral inward in the clockwise sense, as may be seen by writing the differential equation in the form

$$\left(\ddot{q} - \frac{At}{S} - \frac{B}{S} + \frac{RA}{S^2} \right) + \frac{R}{S} \left(\dot{q} - \frac{A}{S} \right) = - \frac{L}{\sqrt{1-\dot{q}^2}} \dot{q}$$

Consider the signs of \dot{q} and $(\dot{q} - A/S)$ in the four quadrants about the equilibrium position. In I for example $(\dot{q} - A/S) > 0$ and the X coordinate of a solution must increase with t ; $\ddot{q} < 0$ so \dot{q} must decrease, giving a clockwise sense to the motion. Similarly the other quadrants may be verified. Some of the solutions starting outside of C will hit one of the lines, but the solution will still be stable. It is easy to show, by a study of the signs of the variables and their rates that a solution can only hit the upper line to the left of the point P_1 with coordinates $X = \frac{R}{S} \left(\frac{A}{S} - K \right)$ and $Y = K$, and that if one does, it will move along the line to the right until it reaches P_1 and then return to the original equation. A similar situation holds for the lower line. If we should start a solution on the upper line to the right of P_1 it would leave the line immediately. The solution is always horizontal (i.e. $\ddot{q} = 0$) on the line through P_1 , the equilibrium point and P_2 .

If $R = 0$ the function Q is constant since $\frac{dQ}{dt} = 0$ and therefore the solutions of the equation

$$Sq + \frac{L}{\sqrt{1-\dot{q}^2}} \ddot{q} = At + B$$

are the equilevel curves in Figure 6.

I have attempted in several different ways to generalize this proof for arbitrary input functions $e(t)$, but so far have no completely rigorous proof. However, some of the arguments come so near as to make me almost certain of complete stability. It can be shown, for example, that two different solutions with the same $e(t)$ cannot definitely diverge; i.e. $|q_1 - q_2| + |\dot{q}_1 - \dot{q}_2|$ cannot become and remain greater than some positive constant (assuming e and e' bounded). Also if two solutions get close together (with respect to both q and \dot{q}), they will certainly converge.

The Effect of Backlash

In order to understand how backlash can cause oscillation, let us first consider a much simplified case. Suppose we have a second order linear system which is less than critically damped with no backlash (Figure 7).

$$Sq + R\dot{q} + L\ddot{q} = e$$

If, at $t = 0$ we suddenly impress $e = E$ (constant) on the system ($q = \dot{q} = 0$), the response is a damped oscillation (Figure 8).

Now in the mechanical system there are only two driven shafts, A and B, and backlash only affects (i.e. directly) the operation of these. Probably the greatest amount is in the adder system driving shaft A. Let us assume for a moment that this is the only backlash present and that its action is as follows: When shaft A reverses direction (i.e. when $\dot{q} = 0$) there is a short pause with the ball carriage stationary while the gear pressures reverse and the gears "catch hold" again in the opposite direction. Suppose this backlash amounts to B_1 turns as measured from the e shaft. It is obvious that the response of the system with backlash is the same as the response would be if there were no backlash and at the time when \dot{q} reverses (previously decreasing - about to increase) we turn the e shaft $-B_1$ turns in such a way as to keep \dot{q} constant while doing this turning.

Similarly at the next reversal we give e a positive increment B_1 keeping \dot{q} constant through this period of backlash.

In other words, the response is that of a system without backlash on which we impress as forcing function a wave which is about as shown in Figure 9.

If the periods of backlash are comparatively short, the small connecting portions (actually quadratic polynomials in time) will have little effect on the response. That is, we can assume a square topped wave with little error in \dot{q} or q especially, due to the smoothing operation of the integrators (or, said another way, due to the high impedance of the circuit to high frequencies).

Now suppose that there is a certain amount of backlash in shaft B. The action of this is to cause the carriage of the upper integrator to remain stationary for a small period when $\ddot{q} = 0$. The same effect would be achieved if, at this time, we suddenly impressed on e a pulse which held the lower integrator at zero and kept changing e at such a rate as to keep the lower integrator there. We keep the integrator at zero long enough so that its output would have turned an amount equal to the backlash in B and then suddenly return it to its proper value. This means that the area of the pulse must equal the backlash. The shape of this pulse would be a linear function of time, but here again it is not highly significant.

The entire system may thus be replaced by one which is free of backlash and subject to a driving function of the type shown in Figure 10, where B_1 is the backlash in A as measured

from e and B_2 is the amount in B as measured from e (in the sense that if e covers an area B_2 , shaft B moves an amount equal to its backlash).

It is easy to see from our diagram that this forcing function is in the correct phase to sustain the oscillation of decay.

The fundamental component of this forcing function is easily found. We have

$$A_1 = \frac{4}{T} \int_0^{\frac{T}{2}} e \sin \frac{2\pi t}{T} dt$$

e may be split into a sum - one term for the square B_1 wave and one for the pulse-like B_2 part. The B_2 pulse is all concentrated near the center of the sine wave where it is nearly unity. Hence approximately

$$\begin{aligned} A_1 &= \frac{4}{T} \int_0^{\frac{T}{2}} \frac{B_1}{2} \sin \frac{2\pi t}{T} dt \frac{4B_2}{T} \\ &= \frac{2B_1}{\pi} 4 f_0 B_2 \end{aligned}$$

The period T of this oscillation is the natural damped period of the system, to within a small error of size comparable to the length of time during which backlash is effective. Hence its

frequency is approximately

$$f_0 = \frac{1}{2\pi} \sqrt{\frac{1}{L_D C} \frac{R^2}{4L^2}}$$

and the magnitude of the fundamental component of the response \dot{q} is

$$I = \frac{\frac{2B_1}{\pi} \quad 4 f_0 B_2}{\sqrt{R^2 \left(\omega_0 L_D - \frac{1}{\omega_0 C} \right)^2}}$$

Providing the quantity $\frac{2B_1}{\pi} \quad 4 f_0 B_2$ is small, the deflection mechanism will behave linearly about its equilibrium position and the above formulae would approximately hold. If $|\dot{q}| \neq 0$ the equilibrium value of inductance $\frac{L}{\sqrt{1-\dot{q}^2/4q}}$ would probably be as good as any to use since the differential inductance is greater on one side and less on the other. At $\dot{q} = 0$ the inductance is greater on each side and a somewhat higher value should be used, depending on $\frac{2B_1}{\pi} \quad 4 f_0 B_2$. If the system is more than critically damped, \dot{q} may or may not have an inflection point depending on the initial conditions. If they are such that the driven shafts do not reverse backlash cannot take effect and there should be no oscillation. However, if they do reverse once, the system may receive the equivalent of a "kick" in such a direction as to cause another reversal and so on, so that oscillation is set up. This problem has not been very well decided but if this happens, the amplitude formula above should still hold, while the frequency formula will not.

The question of "spring backlash" i.e. undesired effects due to elastic deformations of shafts and mounting plates has been raised. According to Hooke's Law the angular strain in a shaft is proportional to the applied torque. This torque in a shaft whose position is $x(t)$ can probably be very well approximated by an equation of the form

$$T = \pm K_1 + K_2 x' + K_3 x''$$

the first term whose sign is that of $-x'$, being due to a coulomb friction load, the second to a viscous friction load and the third an accelerating torque.

It is clear that the coulomb friction term K_1 can be combined with the ordinary gear type backlash treated above, and acts, therefore, like a periodic forcing function. The effect of the other terms is quite different, their presence causes small changes in the parameters λ , μ , and S of the circuit and also adds higher derivatives to the equation. Let us consider only the spring in the shafts feeding $\frac{L}{\sqrt{1-q^2}} \ddot{q}$ (i.e. assume \dot{q} driven rigidly). Evidently

$$\begin{aligned} \frac{L}{\sqrt{1-\dot{q}^2}} \ddot{q} &= (e - a_1 \ddot{e} - a_2 \ddot{e}) \\ &- (Sq - \mu_1 \dot{q} - \mu_2 \ddot{q}) \\ &- (R \dot{q} - \gamma_1 \ddot{q} - \gamma_2 \ddot{q}') \end{aligned}$$

or

$$Sq + (R - \beta_1) \dot{q} - \frac{L}{1 - \dot{q}^2} - \beta_2 - \gamma_1 \ddot{q}$$

$$- \gamma_2 \ddot{q} = (e - a_1 \dot{e} - a_2 \ddot{e}) = e_1(t)$$

Spring in the drive to \dot{q} has a similar effect although complicated by the non-circular sine gears.

If e is a linear function of t , so is e_1 and the forcing function thus contains nothing to create a sustained oscillation. The left-hand side differs only by small quantities from the ideal equation

$$Sq + R\dot{q} - \frac{L}{1 - \dot{q}^2} \ddot{q} = e_1$$

and will therefore surely approach the solution

$$\dot{q} = \frac{\dot{e}_1}{S}$$

Thus we see that the "spring type" of backlash cannot cause sustained oscillation as the "gear" type of backlash can. However, if the gear type is present, the spring type can aid oscillation by reducing the damping. It may be necessary to overdamp in some cases in order to get an effective critical damping.

It should be pointed out that the gear type of backlash may not be quite as simple as we have assumed, particularly in the shafts driving $\frac{L \ddot{q}}{\sqrt{1 - \dot{q}^2}}$. If the integrator carriage load is large compared to the friction loads in the adders and gears, then we are probably justified in assuming that gear pressures in the drive only reverse when the driven shaft reverses. However, if

this is not the case, a backlash effect can easily take place at other times, for example when one of the shafts feeding the adder reverses, without necessarily reversing the driven shaft $\frac{\ddot{q}}{1-\dot{q}^2}$.

The situation could become quite complicated, the equivalent input function containing several different sized steps occurring at different times. However, the fundamental frequency should still be approximately the natural damped frequency of the system, providing the backlash effects are small and occur only during a small fraction of the time.

The fact that backlash can cause a sustained oscillation leads to a criticism of the design of the mechanism, in particular to the method whereby the arcsin function is obtained. Note that reducing the amount of gear backlash $\frac{2B_1}{\pi} 4f_0B_2$ will reduce the amplitude of oscillation proportionately, but apparently the only way to eliminate it completely is to at least critically damp the system for all equilibrium points, so that the shafts do not, in general, reverse direction. In the deflection mechanism as it stands, this would be distinctly disadvantageous, for if we critically damp at the maximum values of $|\dot{q}|$, (the governing points) the system will be much over-damped near $\dot{q} = 0$, and in fact for most values of \dot{q} due to the shape of the inductance curve.

Another related argument against the manner of getting the arcsin is that the response to high frequency error functions depends on the value of \dot{q} . It seems to me that the treatment of error functions should be independent of the target speed -

what is best for one will be best for another - since the prediction error we can tolerate is an absolute quantity, not dependent on the target speed. There may be some objection to this argument on the grounds that at higher target speeds the error function is apt to be larger, and hence the circuit should have a larger impedance, but even so it would only be accidental if the peculiar variation introduced by the sinegear was anything like an approximation to the desired variation.

Finally, a minor argument against the position of the sine gear is that the equation becomes so difficult to handle mathematically. A design of this type must be largely intuitive or experimental - there is not much chance of choosing the constants for the best operation by a mathematical formulation, or of determining to speed of response etc. analytically.

These difficulties might be avoided in several ways. The arcsin might, for example, be introduced as in Figure 11.

No doubt the reason this was not done was because with $|\dot{q}|$ near 1, running the $\sin x$ gear backward is not mechanically practical, the gearing up ratio being too great. This objection could be

overcome in two ways -- either a new gear $K \arcsin x$ to x (k large) could be used and the parameters R , L , S all decreased by a factor of k (or the integrator disks might be speeded up in suitable ratios), or, if this were not mechanically feasible, a rapid response servo mechanism could be introduced in the output, Figure 12.

This system, can, by the way, be solved in closed analytic form when \dot{q} is a constant, and reduced to a quadrature in any case.

The essential feature of this circuit is that the functions of rate finding and smoothing, and of taking the arcsin have been isolated. Each part can be designed to do its own job the best without compromise. It may be noted that the arcsin circuit above also performs a smoothing operation which depends on target speed. By suitable choice of the parameters we can make this large or small as we desire.

The Ideal Rate Finder and Smoother

Let us consider the problem of rate finding and smoothing from a general standpoint, and ask what mathematical operation a machine should perform to act as the "best possible" rate finder. Of course, this question has many answers, depending chiefly on what assumptions we make as to the input function,

and what mathematical limitations we put on the machine. We shall assume throughout that the input function $e(t)$ consists of a series of linear parts with curved connecting portions and with a small superimposed error function, and that we only desire the rate during (that is, some time after the start of) a linear part. In this section we assume there are no limitations whatever on the machine - that we can build a machine to perform any operations we can describe, in particular those a mathematician might use to solve the problem. Now there is considerable experimental and theoretical justification to the theory that the best way to fit a curve of a given type to a set of points subject to an observational error is in the least square sense. If we assume this to be true in our case, and attempt to fit a straight line to the last a seconds before t_1 of the curve $e(t)$, we must minimize the integral

$$I = \int_{t_1-a}^{t_1} e - (At+B)^2 dt$$

with respect to A and B . The quantity a represents the length of the curve used in the fitting process. We would like to use as much of the curve as actually represents a linear segment to get the best accuracy, but certainly no more. A person doing the curve fitting could look at $e(t)$ and see fairly well where the curve showed a real tendency to depart from linearity, and select accordingly. Mathematically it could be done as follows. Suppose the

standard deviation of the error is σ and that errors of more than say 4σ are almost certainly due to a significant departure from linearity in the curve. We could choose a such that it is as large as possible without making the error $|e-(At+B)|$ (A, B chosen to minimize I) $t_1-a \leq t \leq t_1$ greater than 4σ . In other words we use as much of the curve as we can assume linear within observational errors. As a final refinement of the solution it might be desirable to include a weighting function $W(a,t)$ in the integral I , weighting the more recent values more heavily. The final evaluation of the rate is then the value of A given when we minimize the function

$$I(A,B,a) = \int_{t_1-a}^{t_1} [e-(At+B)]^2 W(t,a) dt$$

on A and B , a fixed, giving A and B as functions of a , and then choose a as large as possible with

$$|e-(At+B)| \leq K\sigma \quad t_1-a \leq t \leq t_1$$

This solution can be put into a more explicit form, but even when greatly simplified it appears that it would be quite difficult to carry out the calculations accurately by mechanical means. The main difficulty is that apparently such a machine must be capable of remembering exactly the past history of an arbitrary function, e or something derived from it. The only methods I know of doing this are quite inaccurate, or else very complex, and it seems likely that the gain in mathematical precision of the above

formulation would be more than offset by a loss in mechanical precision.

Differential Analyzer Types of Machines

To become a bit more practical, let us now confine our attention to machines of what might be called the differential analyzer type. By this, we mean machines constructed of a finite combination of adders, integrators, and function elements (e.g. non-circular gears). Two shafts $e(t)$ and kt enter the machine and one shaft $u(t)$ leaves the machine. It can be shown that any such system must satisfy a differential equation of the type

$$f(q, \dot{q} \dots q^{(n)}, t) = e(t)$$

with

$$u(t) = q^{(i)}.$$

First, we ask what can be said about the form of this equation to make the machine act as a satisfactory rate finder in our sense.

1. With the same initial conditions and the same $e(t)$ the machine should certainly respond the same independent of the time of start. Hence f does not depend on t .
2. When $e = 0$ At B the equation must have an equilibrium solution

$$q^{(i)} = A \quad q^{(i-1)} = 0$$

$$q^{(i-1)} = At \quad e$$

338

If $i > 1$, the carriage of an integrator will be continuously moving in the equilibrium condition. This does not seem practical for the initial conditions may be anything depending on past history, and the integrator would surely go off scale in many cases. Obviously from the equilibrium solution, i is not 0, for this would imply a constant equal to a linear function of time. Hence $i = 1$ and $q' = u(t)$.

3. Let

$$f(x, y) = f(x, y, 0, \dots, 0)$$

due to the equilibrium solution

$$f(At + C, A) = At + B$$

for all A, B, t .

$$\frac{\partial f}{\partial t} = \frac{\partial f(x, y)}{\partial x} \quad A = A$$

$$f(x, y) = X + h(y)$$

4. Assuming f is fairly "well behaved", we have near $q = \dot{q} = \dots$

$q^{(n)} = 0$ (i.e. near equilibrium)

$$f = f(q, \dot{q}, 0, 0, \dots, 0)$$

$$= \frac{\partial f}{\partial q} q^{(n)} + \frac{\partial f}{\partial \dot{q}} \dot{q}^{(n)} + \dots + \frac{\partial f}{\partial q^{(n)}} q^{(n)}$$

$$= q + h(\dot{q}) + a_2 \ddot{q} + \dots + a_n q^{(n)}$$

and the differential operation depends on the coefficients $a_2 \dots a_n$ and $h(\dot{q})$. As this differential operation should not depend on t , the a_i must be independent of q , for in equilibrium q changes with t . They may depend on \dot{q} however in which case the differential operation depends on the target speed, which may or may not be desirable. In the deflection mechanism this is the

$$\text{case, } a_2 = \frac{1}{1-\dot{q}^2}$$

5. With \dot{q} near A the above reduces to

$$f = q + a_1 \dot{q} + a_2 \ddot{q} + \dots + a_n q^{(n)} + b$$

where $a_1 = h'(A)$ and $b = h(A) - Ah'(A)$. To eliminate backlash oscillation the roots of this equation should all be real and for stability all should be negative, for all desired A .

6. For complete stability, there are no doubt further requirements on the form of f . This problem, however, is still unsolved.

The above are only requirements on the form of f so that it actually does find a satisfactory rate. To find the best form of f would require a very elaborate mathematical analysis if possible at all.

If we restrict our machine still further and assume a linear differential equation with constant coefficients, it is possible to give a fairly rational analysis leading to the best values of the coefficients. The question is this. Given the equation

$$a_0 q + a_1 q' + \dots + a_n q^{(n)} = e$$

What values of the coefficients $a_0 \dots a_n$ give the best rate-finding smoothing properties? From what we said above, it seems that the characteristic equation

$$\sum a_n p^n = 0$$

should have only real negative roots and that the rate found will be q' . We may normalize the equation by assuming $a_0 = 1$ so that q' is actually the rate and not merely proportional to it. In the Heaviside symbolic notation, we have

$$q' = \frac{p}{(b_n p + 1)} \cdot e$$

writing the polynomial in the factored form. The b_k are positive real numbers and are the time constants in the transient part of the response. We assume the b_k arranged in increasing magnitude.

Let us frame the problem as follows. Keeping the speed of response of the circuit the same, what values of the b_k give the best attenuation of the error function. Of course, the trouble appears in trying to decide what we mean by keeping the speed of response the same. One answer is that we keep the maximum time constant, that is b_n , the same. This may be partially justified on the following grounds: 1. For "almost all" initial conditions, the term $A_n e^{-\frac{t}{b_n}}$ will eventually dominate the transient response,

the other terms becoming arbitrarily small in comparison. The only time when this fails is when the coefficient A_n happens to come out zero.

2. In the worst cases (other coefficients small in comparison) the b_n term dominates for all t , and the machine should perhaps be designed with the worst conditions as governing.

3. If we use this criterion, it is easy to show that for best attenuation of error frequencies all the b_k should be equal. For the magnitude of the transfer admittance (e to q') is

$$Y(j) = \frac{e}{V(1 - b_k^2 \omega^2)}$$

which is obviously smallest when each b_k is made as large as possible, for all frequencies. That is, each $b_k = b_n$ the maximum.

Another way the "same speed of response" might be interpreted is in terms of the expected area under the transient time curve. Keeping the standard deviation of this area constant seems to give the same evaluation of the b_k as above but there are certain statistical assumptions in my proof that may render it invalid.

If the characteristic equation has real roots, it may be set up nicely as in Figure 13.

This circuit appears to have an advantage from the backlash point of view over the more obvious one shown in Figure 14.

It seems quite possible, however, that the use of nonlinear equations could offer a real advantage. Consider the equation

$$S(\dot{q}) \dot{q} + R(\dot{q}) \ddot{q} + L(\dot{q}) \dot{q} = 0$$

where the three coefficients are functions of \dot{q} . When the system is at equilibrium, it acts approximately like:

$$S(0) \dot{q} + R(0) \ddot{q} + L(0) \dot{q} = 0$$

and these three constants could be adjusted to give critical damping and a good attenuation of the error function frequencies. On the other hand, when we are not at or near equilibrium, \dot{q} is (usually) considerably different from zero. The values of the three coefficients could be adjusted in this case to give a very rapid response, and thus approach the equilibrium position faster. It is possible, however, that there is some fundamental error in this reasoning, for example, that an attempt to do this would necessarily cause oscillation.

Recommended Linear Circuits.

Although, as has been indicated, there is much possibility for further research, on the basis of my present knowledge, I would recommend the circuits in Figure 15, for rate-finding and smoothing.

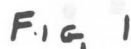
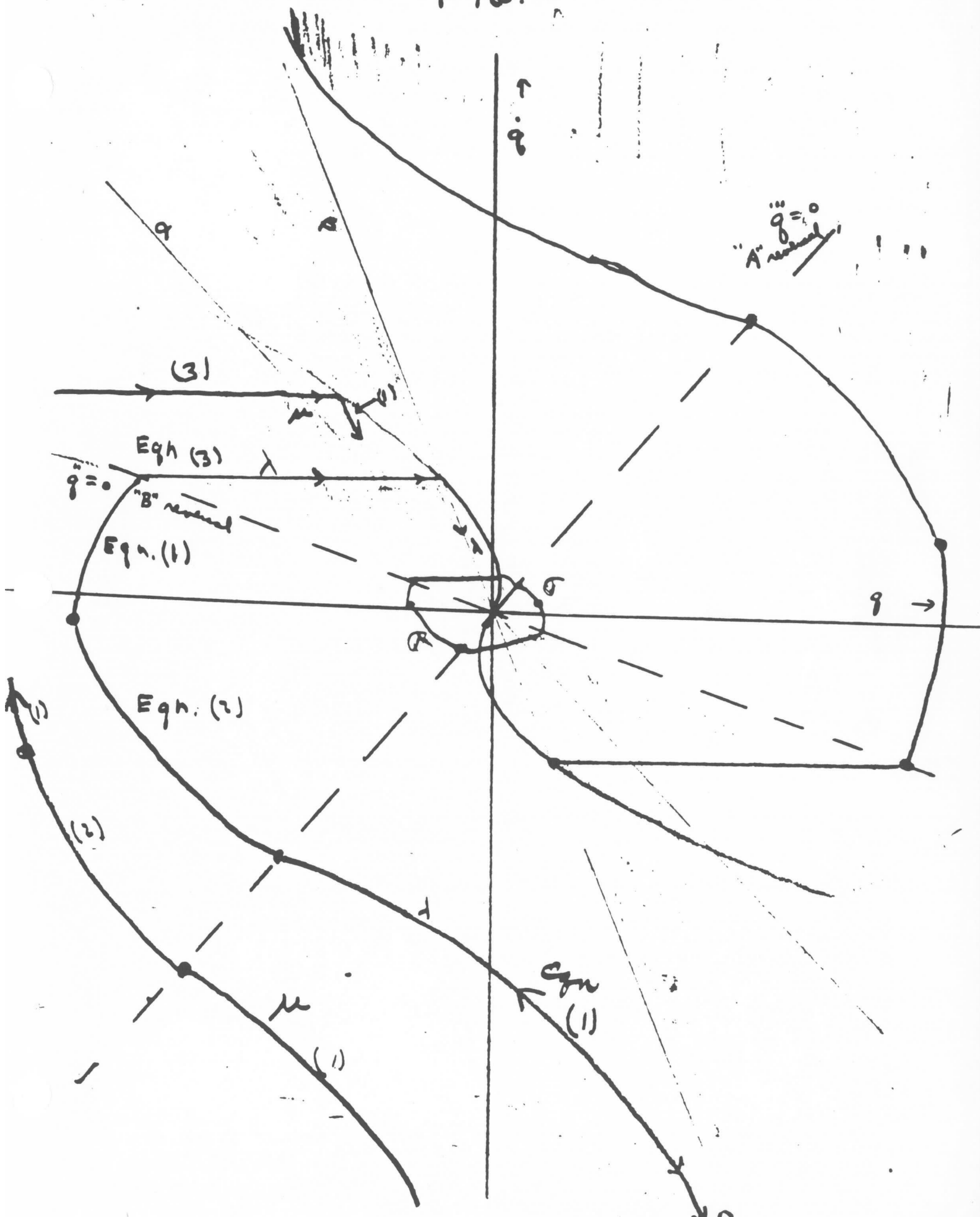


Fig. 3



[9]

20-105

A HEIGHT DATA SMOOTHING MECHANISM

Claude E. Shannon

5/26/41

①
1/2/56

misc.

A HEIGHT DATA SMOOTHING MECHANISM

The schematic diagram of a new type of height data smoothing mechanism is shown in Figure 1. The discontinuous height data $e(t)$ is fed into the input shaft at intervals. This drives a differential, connected also to the ball carriage and roller of an integrator whose disk is turned by a constant speed motor. A correcting handwheel and the integrator roller feed another differential whose output is the output of the device. The output and input of the machine are compared through a differential feeding dial. The operator is supposed to turn the handwheel in such a way that the positive and negative oscillations of the dial about zero are equal.

The actual height of the target $h(t)$ is a continuous function of time and we may assume that just after each reading $e(t)$ is an approximation to this. Thus $h(t)$ and $e(t)$ might be as shown in Figure 2.

The shaft $y(t)$ clearly satisfies the equation

$$(1) \quad y + \frac{1}{\alpha} y' = e(t) \quad .$$

The x shaft satisfies

$$(2) \quad x(t) = y(t) + e(t)$$

and the dial reads

$$(3) \quad D(t) = e(t) - x(t) \quad .$$

During the period between height readings the position of the $e(t)$ shaft is constant, say $e(t_n)$, the reading taken at t_n .

$$y + \frac{1}{m} y' = e(t_n)$$

$$y = e(t_n) + A_n e^{-m(t-t_n)} \quad t_n = t < t_{n+1} \quad .$$

Since y is obviously continuous, it will follow a curve consisting of a series of connected exponentials, each with the same time constant, $\frac{1}{m}$. The continuity of the curve implies

$$e(t_n) + A_n e^{-m(t_{n+1}-t_n)} = e(t_{n+1}) + A_{n+1} e^{-m \cdot 0}$$

$$A_{n+1} = A_n e^{-m(t_{n+1}-t_n)} - e(t_{n+1}) + e(t_n) \quad .$$

Assuming the intervals between readings the same, say a seconds, the response y for two different time constants $m_1 a = \ln 2$ and $m_2 a = \ln 10$ are shown in Figure 3.

The larger the time constant, the more the lag in response of $y(t)$, but the smoother the curve. This may be seen another way: the e to y system is equivalent to an R , L circuit with position of shafts analogous to voltage as shown

in Figure 4. With $\frac{L}{R}$ small y follows e closely including the irregularities. With $\frac{L}{R}$ large $y(t)$ is smooth compared to e but lags considerably.

Movement of the handwheel does not affect $y(t)$ but shifts $x(t)$ up or down with respect to y . If the operator turns the wheel to give equal positive and negative movements of the dial, it may be seen that in the "steady state" (say with $f(t) = at$) there is a constant lag even when the damping is low and the interpolation nearly linear. In this case the system bridges linearly between the mid-ordinates of the steps, while actually it should bridge between the points $(t_n + 0)$. With higher damping the shape becomes worse but the interpolated exponentials are nearer to the true curve most of the time. We shall find a formula for the best time constant of the system under the following assumptions

1. That the "best" time constant is the one making the actual error least in the mean square sense.
2. That we may take as the true curve, so far as our knowledge goes, the linear interpolation between the points $t_n + 0$. This may be justified by the fact that the device cannot in any way perform higher order interpolation - the curve $y(t)$ is convex upward whenever $e(t)$ increased in its last step over the final value of y from the preceding step, and this is quite independent of the curvature of $e(t)$.

3. That the system is in a "steady state", that is, that in the step under consideration $y(t)$ ends at the same distance below $e(t)$ as it was just before the step.

4. That the steps come at approximately equal intervals of a seconds.

An interval under these conditions is shown in Figure 5. Here we assumed that the handwheel was turned to give a ratio of $\frac{c}{b-c}$ as deflection of the dial just after to just before a step.

We have

$$y = A e^{-mt}$$

with

$$y(0) - b = y(a)$$

1. That the "best" time constant is the one making the

$$A - b = A e^{-am}$$

actual error least in the mean square sense.

2. That we may take as the true curve, so far as our

$$A = \frac{b}{1 - e^{-am}}$$

knowledge of the linear interpolation between the points $t = 0$ and $t = a$. This may be justified by the

Hence

fact that the device cannot in any way perform

$$y = \frac{b e^{-mt}}{1 - e^{-am}}$$

higher order than $y(t)$ is one-
way upward whenever $e(t)$ increased in its last step

over the final value of y from the preceding step,

and this is the basis of the hypothesis of

$$x = y - y(0) + c$$

$$= -b \frac{1 - e^{-mt}}{1 - e^{-am}} + c$$

The integral of the squared error per second is then

$$I^2 = \frac{1}{a} \int_0^a \left[-b \frac{1 - e^{-mt}}{1 - e^{-am}} + 0 + \frac{b}{a} t \right]^2 dt$$

$$\frac{I^2}{b^2} = \frac{1}{a} \int_0^a \left[k + \frac{t}{a} - \frac{1 - e^{-mt}}{1 - e^{-am}} \right]^2 dt$$

where $k = \frac{c}{b}$

letting $D = ma$, $u = \frac{t}{a}$

$$\left(\frac{I}{b}\right)^2 = \frac{1}{a} \int_0^1 \left[k + u - \frac{1 - e^{-Du}}{1 - e^{-D}} \right]^2 du$$

$$= \frac{1}{a} \int_0^1 \left[k^2 + u^2 + \frac{(1 - e^{-Du})^2}{(1 - e^{-D})^2} + 2ku \right. \\ \left. - 2k \frac{1 - e^{-Du}}{1 - e^{-D}} - 2u \frac{1 - e^{-Du}}{1 - e^{-D}} \right] du$$

$$= \frac{1}{a} \left[k^2 u + \frac{u^3}{3} + \frac{1}{(1 - e^{-D})^2} \left(u + \frac{2}{D} e^{-Du} - \frac{e^{-2Du}}{2D} \right) \right]$$

$$\begin{aligned}
 & + k u^2 - \frac{2k}{1 - e^{-D}} \left(u + \frac{1}{D} e^{-Du} \right) \\
 & - \frac{2}{1 - e^{-D}} \left[\frac{u^2}{2} + e^{-Du} \left(\frac{u}{D} + \frac{1}{D^2} \right) \right] \Bigg|_0^1 \\
 & = \frac{1}{a} \left[k^2 + \frac{1}{3} + \frac{1}{(1 - e^{-D})^2} \left[1 - \frac{2}{D} (1 - e^{-D}) + \frac{1 - e^{-2D}}{2D} \right] \right. \\
 & \left. + k - \frac{2k}{1 - e^{-D}} \left(1 - \frac{1 - e^{-D}}{D} \right) - \frac{2}{1 - e^{-D}} \left[\frac{1}{2} + e^{-D} \left(\frac{1}{D} + \frac{1}{D^2} \right) - \frac{1}{D^2} \right] \right] \\
 & = \frac{1}{a} \left[\left(\frac{1}{3} + k + k^2 \right) + \frac{2k}{D} + \frac{2}{D^2} + \frac{1}{(1 - e^{-D})^2} - \frac{(2 + 4k) \cdot D + 3 + 3e^{-D}}{2D (1 - e^{-D})} \right].
 \end{aligned}$$

It is evident from physical considerations that the minimum of this expression occurs for a fairly large D . In fact the error curve was plotted for $k = .5$ (Figure 6) and the minimum is seen to be at about 7 or 8. With D this large the above expression is very nearly equal to

$$\phi = \frac{1}{a} \left[\left(\frac{1}{3} + k + k^2 \right) + \frac{2k}{D} + \frac{2}{D^2} + 1 - \frac{(2 + 4k) \cdot D + 3}{2 D} \right]$$

since e^{-D} is very small. To locate the minimum we have

$$\frac{d\phi}{dD} = -\frac{2k}{D^2} - \frac{4}{D^3} - \frac{2D(2 + 2k) - 2[(2 + 4k)D + 3]}{4 D^2} = 0$$

whence

$$(6 - 8k) D = 16$$

$$D = \frac{8}{3 - 4k}$$

For $k = \frac{1}{2}$

$$D = 8$$

Since the minimum is so flat (Figure 6) this formula is certainly close enough. However a second approximation may be found as follows: for x small $\frac{1}{1-x} \approx 1 + x$. Using this in the exact expression to eliminate the denominators we get as a second approximation

$$\psi = \frac{1}{2} \left[\left(\frac{1}{3} + k + k^2 \right) + \frac{2k}{D} + \frac{2}{D^2} + (1 + 2e^{-D}) - (1+k)(1+e^{-D}) - \frac{3}{2D}(1+e^{-D}) - \frac{3}{2D}(1+e^{-D})e^{-D} \right]$$

$$\frac{dy}{dD} = 0 = -8 + (3-4k)D + [6D(D+1) + 2D^3(k-1)]e^{-D} + 6D(D+1)e^{-2D}$$

Using the first approximation to obtain the values involving exponentials, a better value may be obtained. For $k = \frac{1}{2}$ the second approximation is $D = 8.03$. The first and second approximations are plotted in Figure 7.

With $k = \frac{1}{2}$ the curve $x(t)$ is plotted for an interval with the "best" D , in Figure 8. It will be noted that the curve is highly damped in comparison to the time between readings. The RMS error is then equal to

$$\frac{I}{b} = \sqrt{\frac{.053}{a}} = \frac{.23}{\sqrt{a}}$$

It is interesting to compare this with the RMS errors obtained under other conditions. If the device is not used at all, but a direct coupling made between the input and output, the RMS error between the step function and the linear interpolation between points $t_n + 0$ is

$$\left(\frac{I}{b}\right)^2 = \frac{1}{a} \int_0^a \left[0 - \left(-\frac{t}{a}\right)^2\right] dt$$

$$\frac{I}{b} = \frac{1}{\sqrt{3a}} = \frac{.577}{\sqrt{a}}$$

so that the RMS error has been reduced to 40% of this value.

In Figure 9, the output of the smoothing mechanism, $x(t)$, is plotted for a certain forcing function $e(t)$, using the "best" value of m . It may appear that the output is still far from smooth, and this is in a sense true, but it must be remembered that the variations in $e(t)$ are here greatly exaggerated over what would be expected in practice.

Finally it should be pointed out that a very material improvement in operation could be obtained if the operator were trained to turn the handwheel to obtain a ratio $\frac{a}{b}$ nearer to zero than $\frac{1}{2}$. This, however, would probably be impractical.



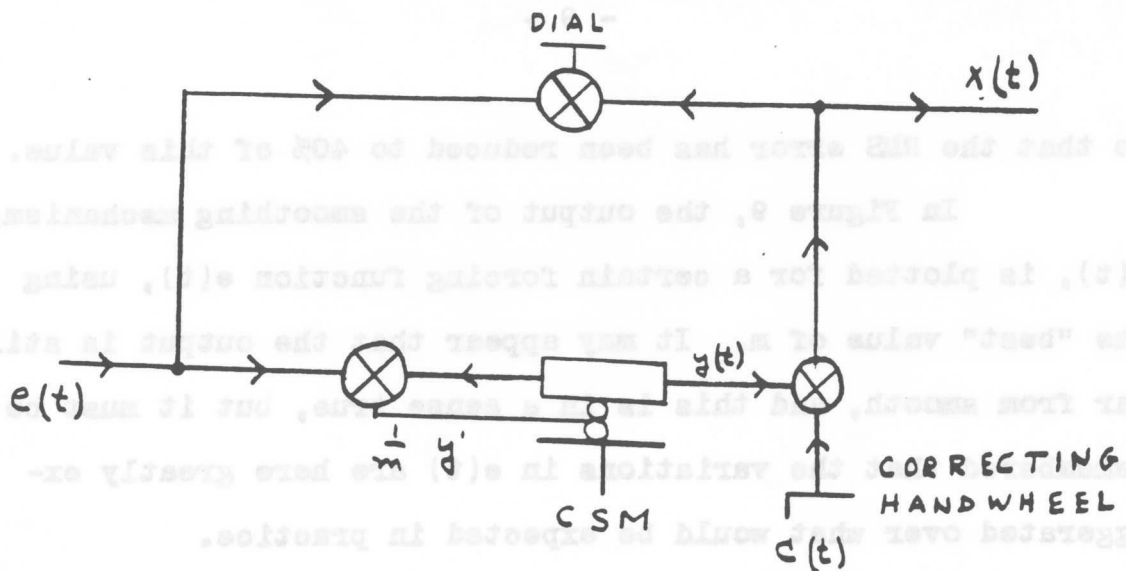


FIG. 1

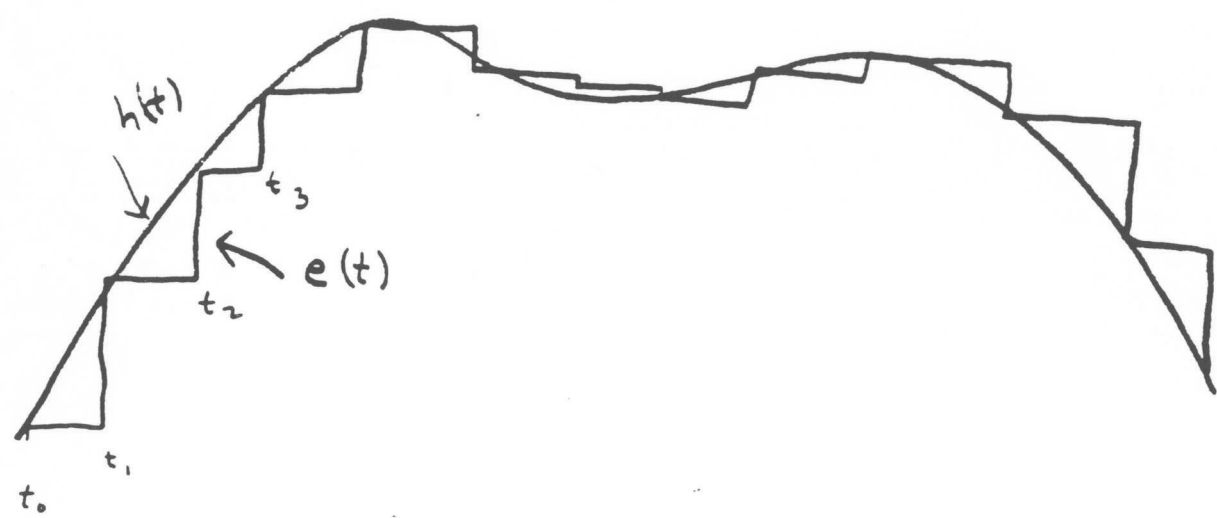


FIG. 2.

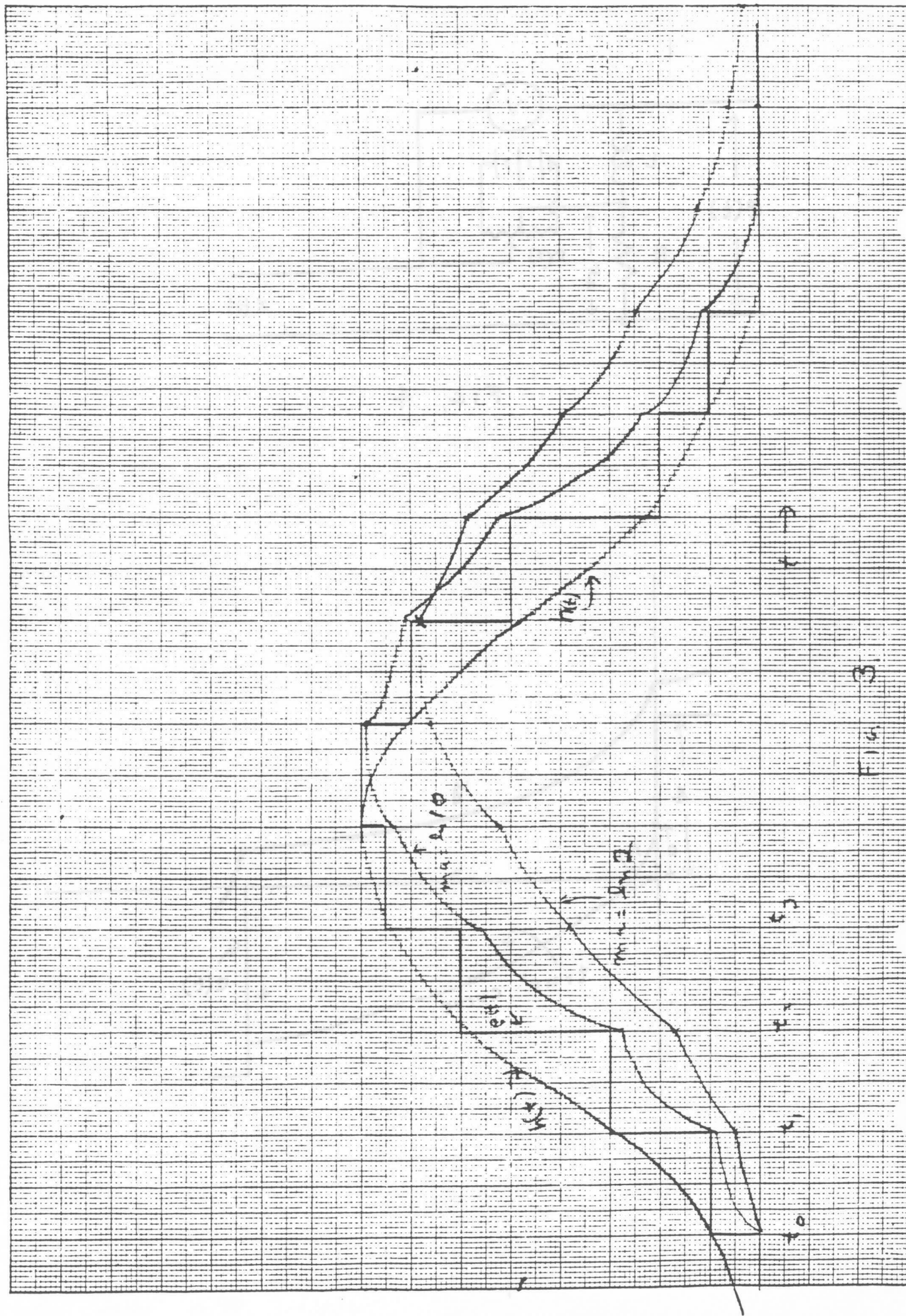


Fig. 3

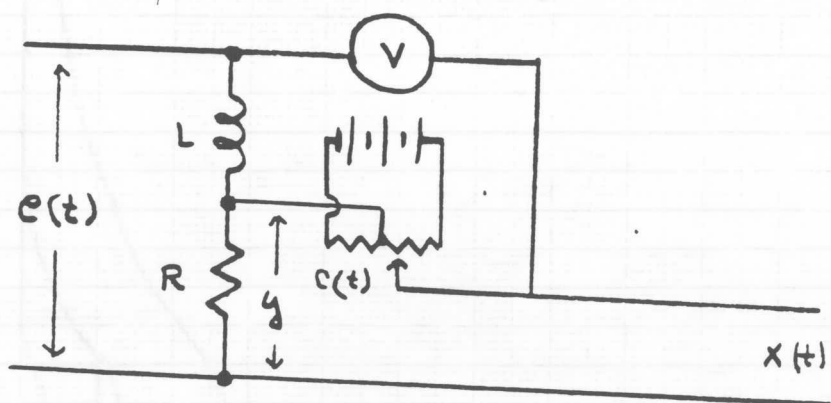
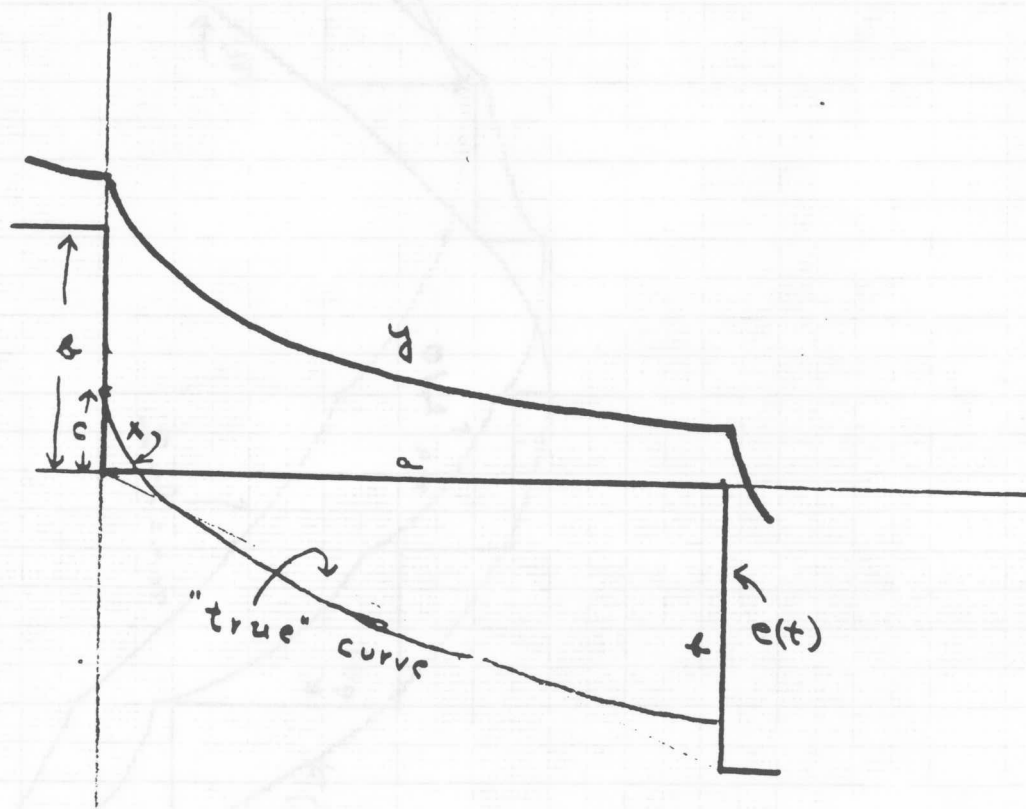


FIG. 4



$K = 1.5$

VARIATION OF ERROR
WITH $D \propto m$

ASYMPTOTIC VALUE AS $D \rightarrow \infty$



Fig. 6

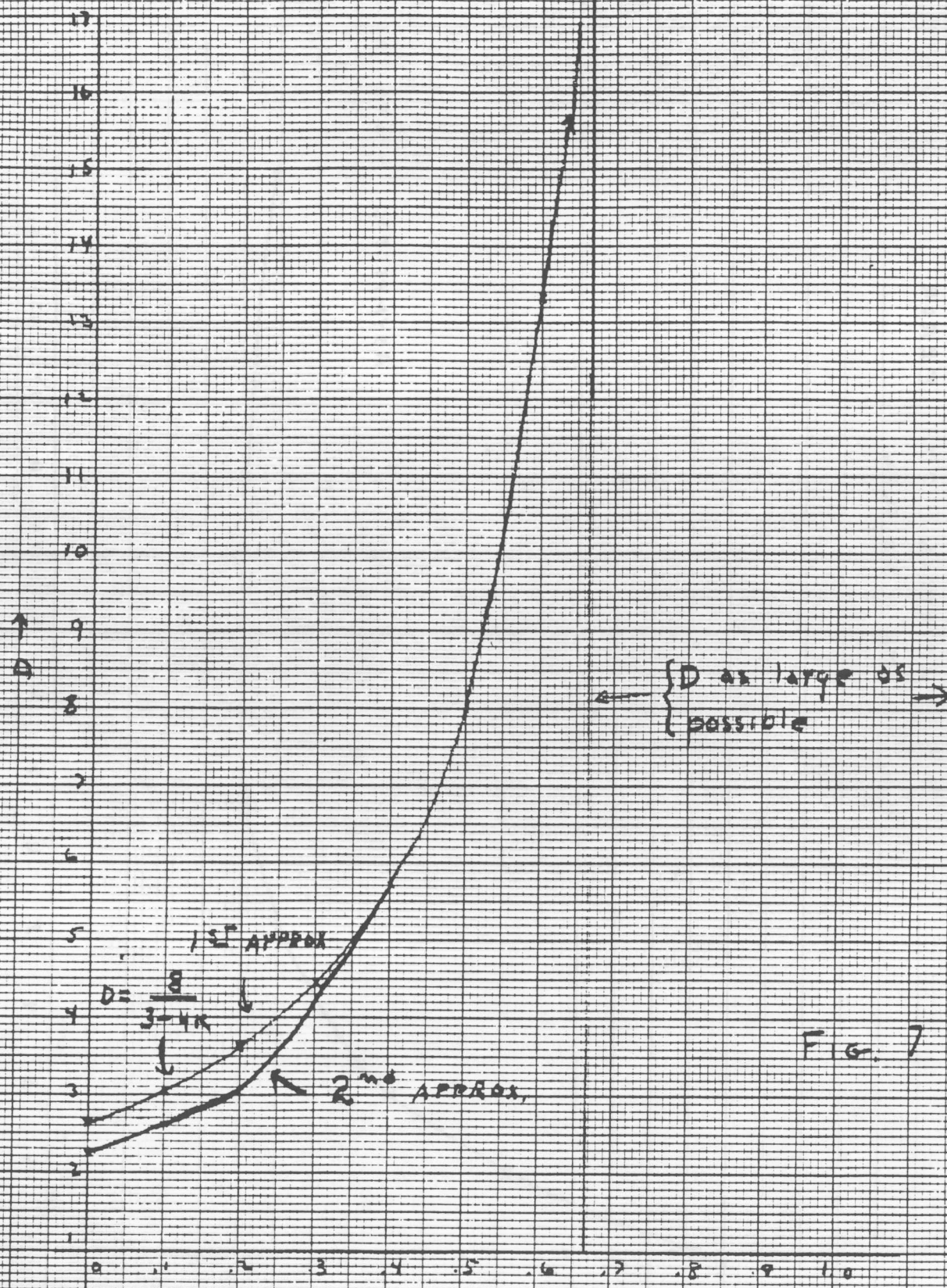


FIG. 7

Value of $D = ma$ to minimize squared error
vs. $K = \frac{c}{b}$

FIG. 8.
BEST CURVE
 $K = 5, D = 8$

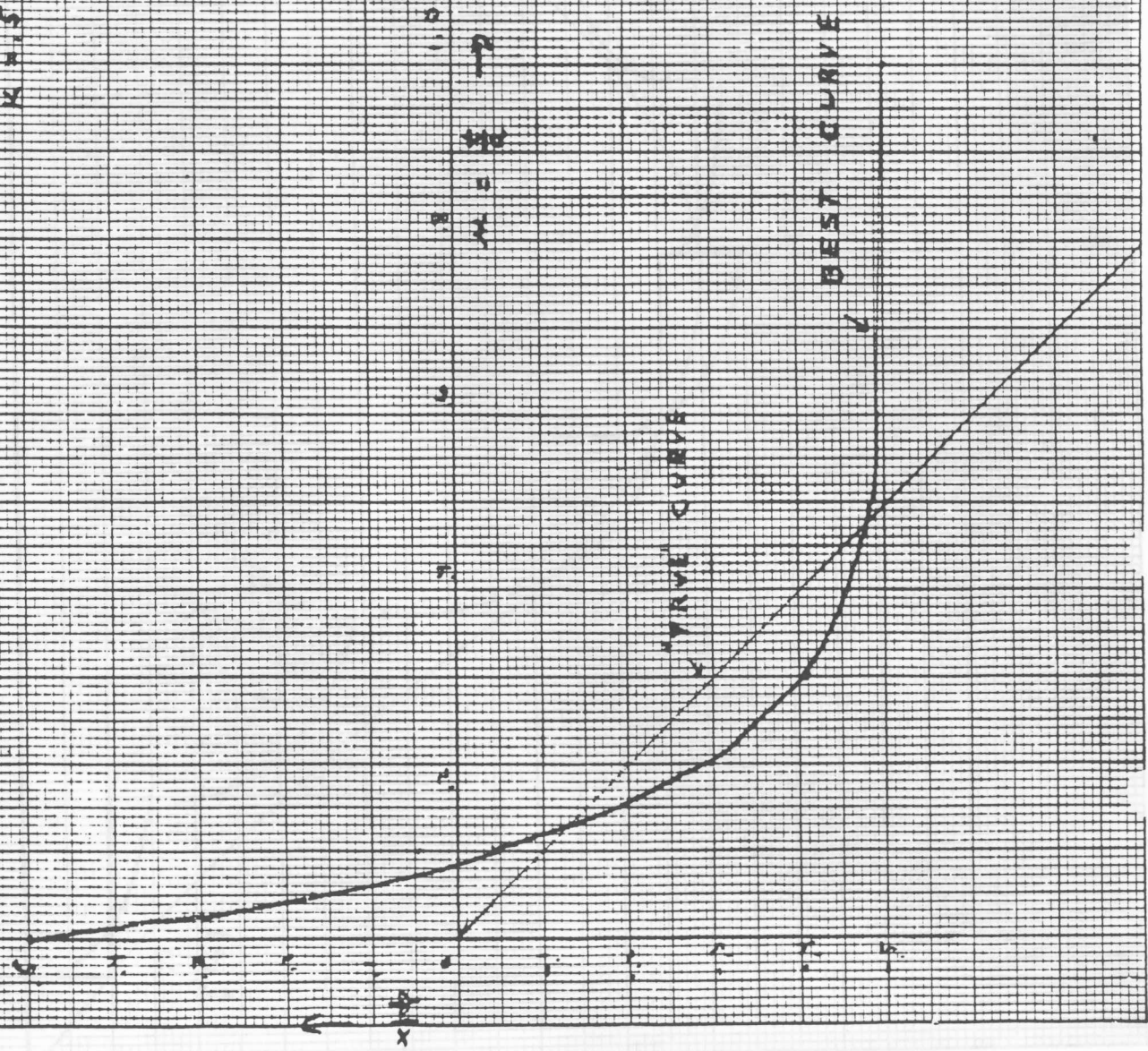
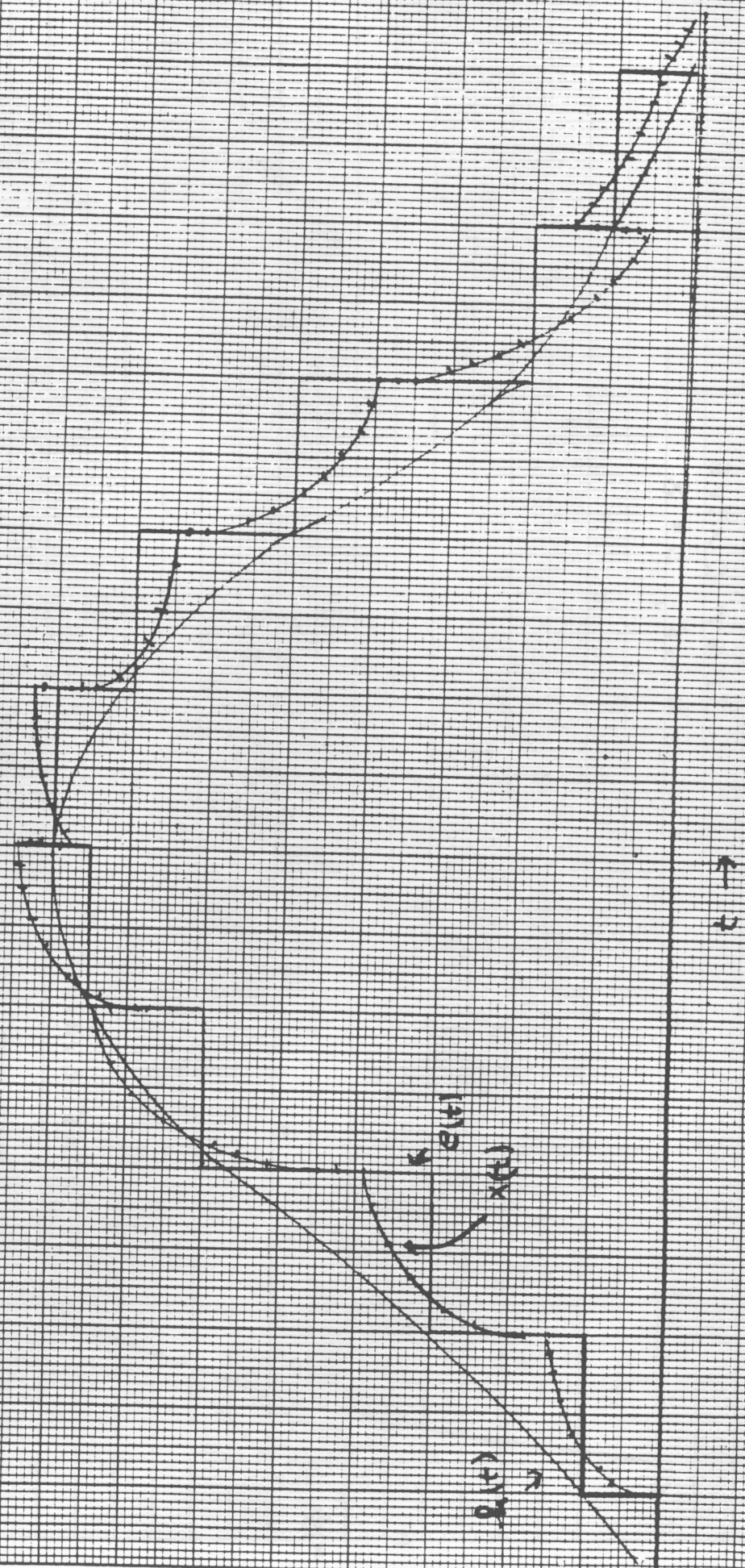


Fig. 9.



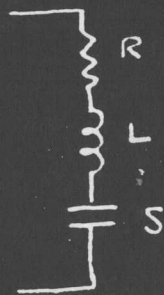
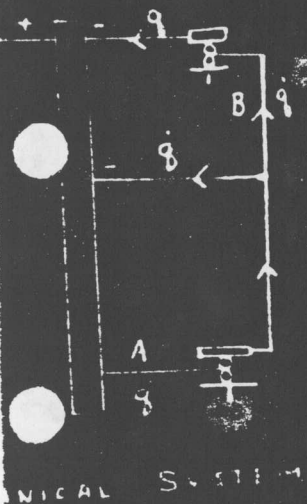
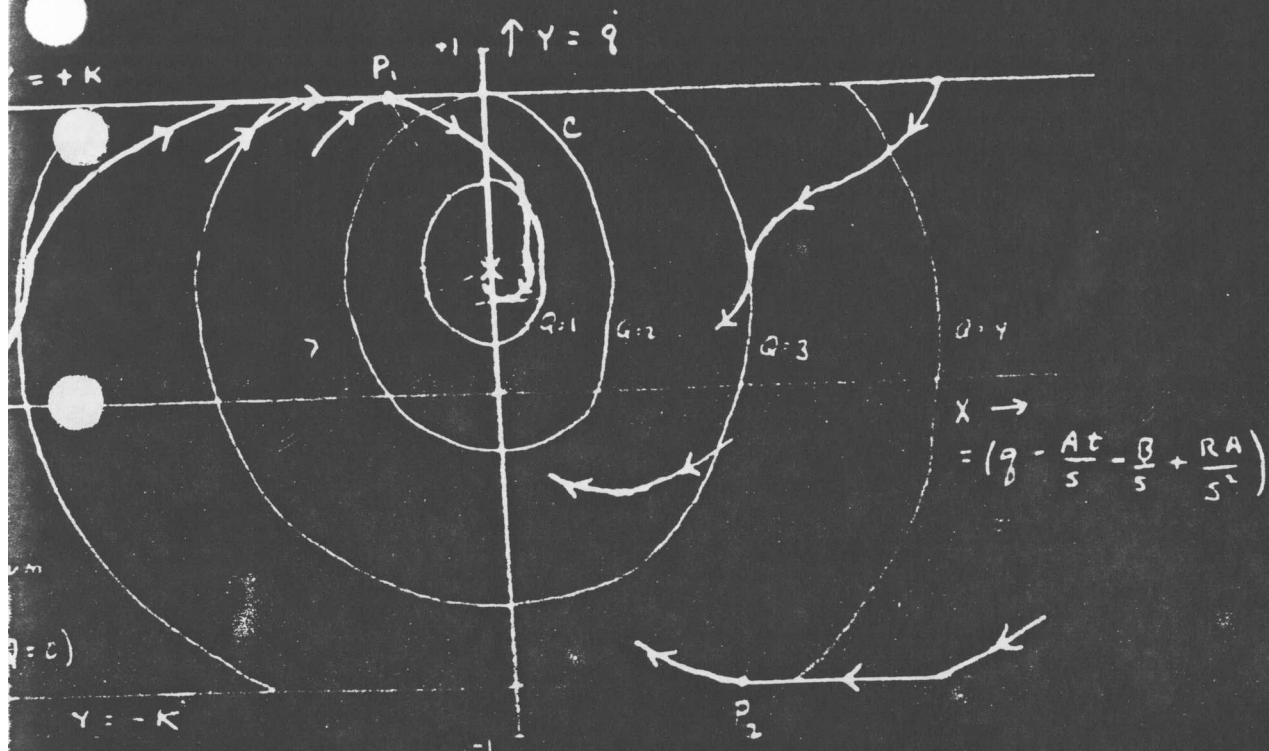


Fig. 7.

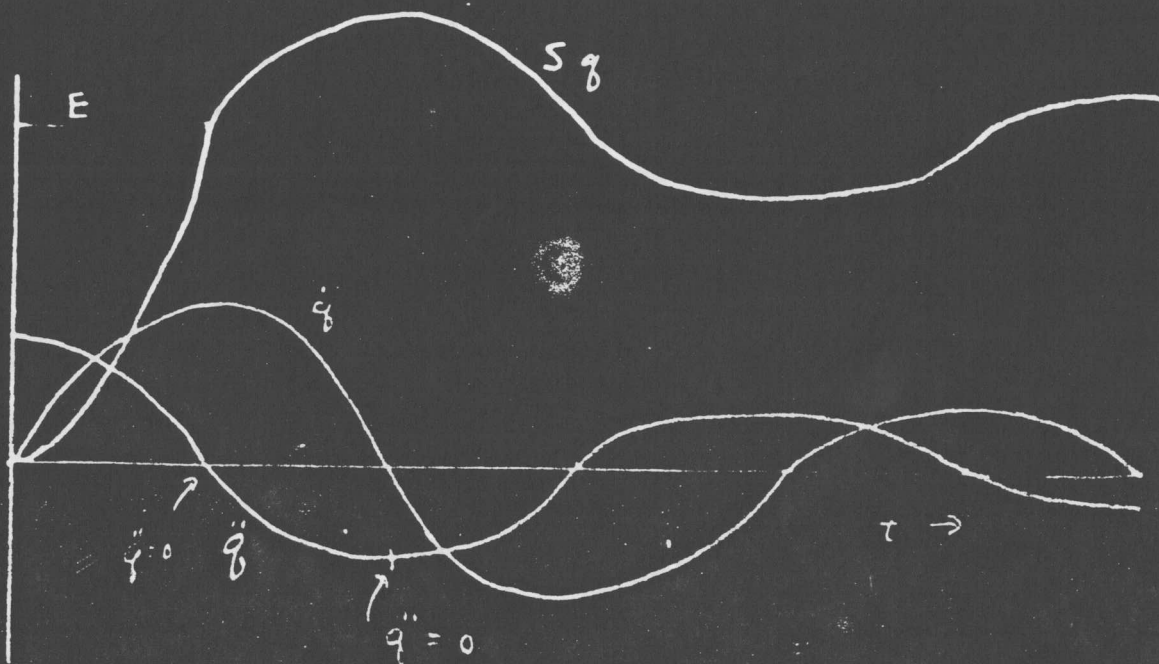


FIG 8.

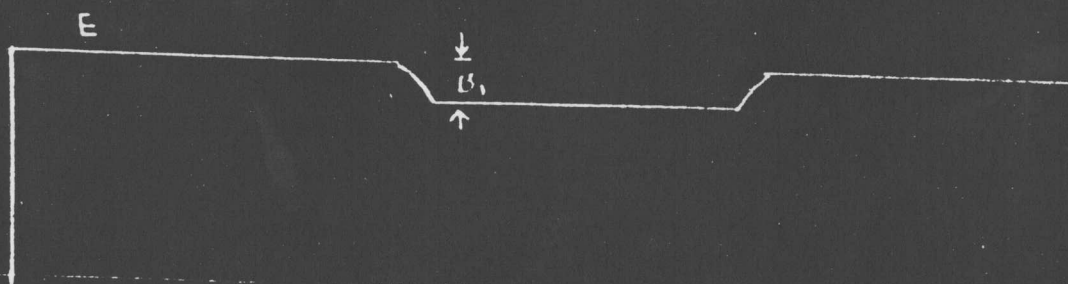


FIG 9

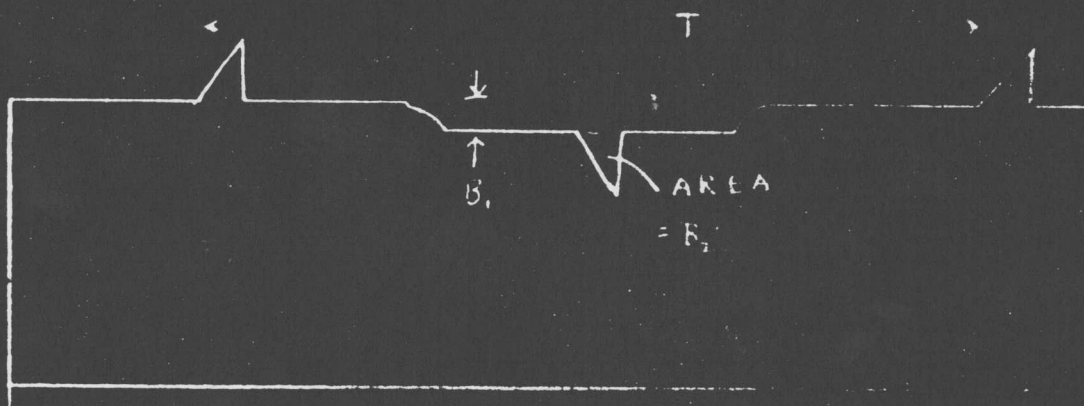


FIG. 10.

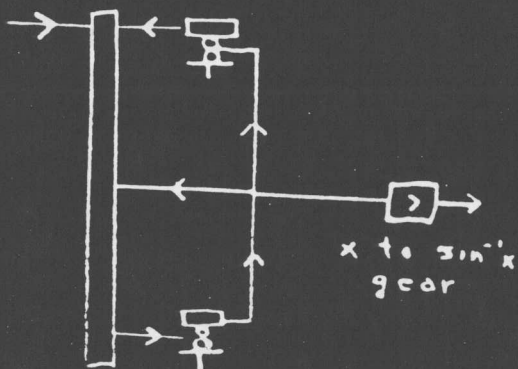


FIG. 11

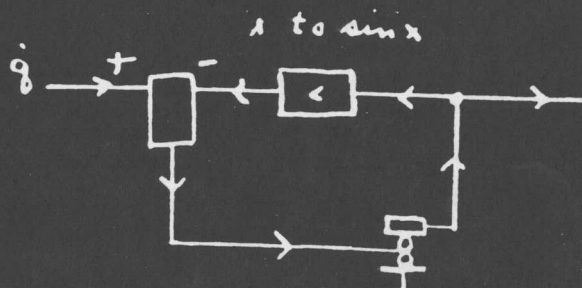


FIG. 12

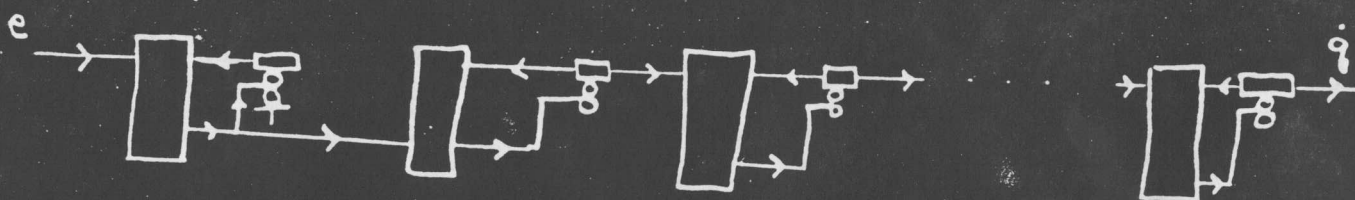
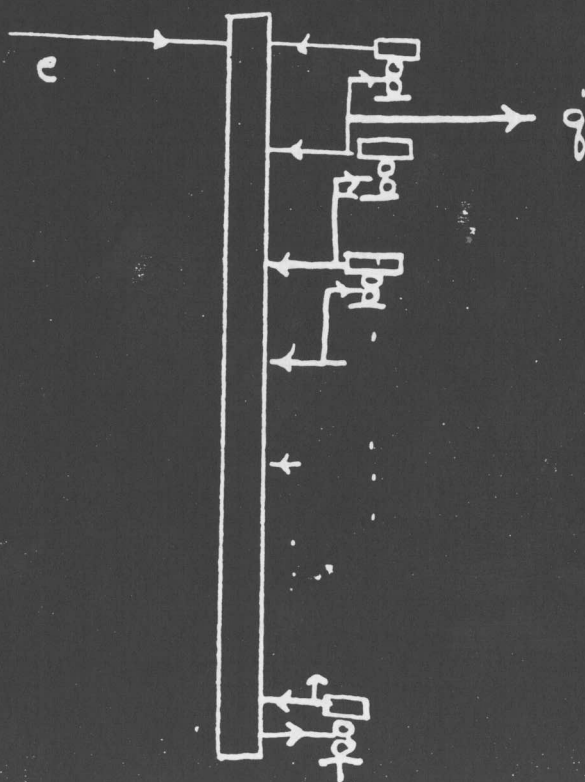
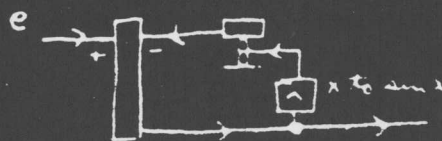


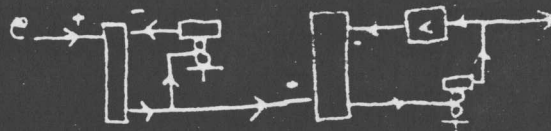
FIG. 13



1 INTEGRATOR



2 INTEGRATORS



3 OR MORE INTEGRATORS

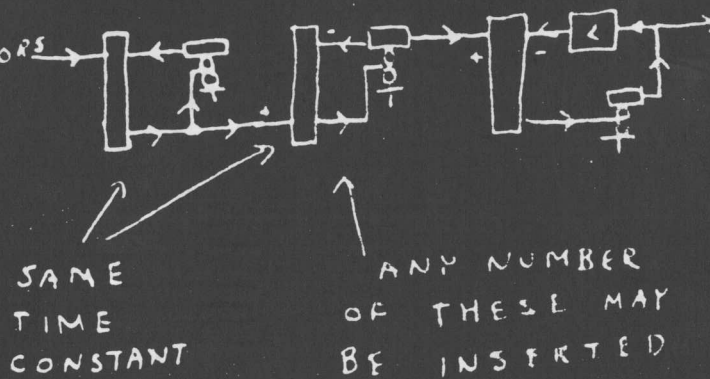


FIG. 15

[11]

Some Experimental Results in the Deflection Mechanism

In a previous report, "A Study of the Deflection Mechanism and Some Results on Rate Finders," a mathematical study was made of a new type of deflection mechanism. The present paper is a further study of this mechanism and a report on some experimental results obtained on the differential analyzer.

For convenience in reference, the schematic diagram of the machine is repeated in Fig. 1. In the report mentioned, the utility of the middle part of the device was questioned. This arose from a misunderstanding of the basic assumptions underlying the design and was cleared up in a conference with Dr. Tappert. The writer's analysis was under the assumption that the mechanism was designed to find rates for linear forcing functions only (i.e., that higher order terms were small by comparison), and the analysis is still valid if this is true. However, in practice, it appears necessary to give the correct steady state rate (except for the non-linearity of the sine gear) for an arbitrary

SOME EXPERIMENTAL RESULTS ON THE DEFLECTION MECHANISM

A schematic diagram of the machine is repeated in Fig. 1. In the report mentioned, the utility of the middle part of the device was questioned. This arose from a misunderstanding of the basic assumptions underlying the design and was cleared up in a conference with Dr. Tappert. The writer's analysis was under the assumption that the mechanism was designed to find rates for linear forcing functions only (i.e., that higher order terms were small by comparison), and the analysis is still valid if this is true. However, in practice, it appears necessary to give the correct steady state rate (except for the non-linearity of the sine gear) for an arbitrary characteristic forcing function. Actually the middle part (often referred to hereafter as the "x" part) of the device is certainly well worth while, as will be seen from some of our experimental curves.

If a linear mechanism has a transfer admittance $Y(j\omega)$ from input $e(t)$ to output $d(t)$ then

$$d(j\omega) = Y(j\omega)E(j\omega)$$

where E and d are the transforms of e and d . It is easily seen from transform theory that if $e(t) = at + b$, a necessary and sufficient condition that $d(t) \rightarrow a$ as $t \rightarrow \infty$ is that

$$Y(0) = \lim_{s \rightarrow 0} \frac{dY}{ds} = 1$$

If this condition is satisfied the system may be called a first order rate finder -- after the transient has died out, the output is the derivative of the input whenever the input is a first order function.

June 26, 1941

$$Y(0) = 0 \quad Y'(0) = 1 \quad Y''(0) = 0 \quad Y'''(0) = 0 \quad \dots$$

Some Experimental Results on the Deflection Mechanism

In a previous report, "A Study of the Deflection Mechanism and Some Results on Rate Finders," a mathematical study was made of a new type of deflection mechanism. The present paper is a further study of this device and a report on some experimental results obtained on the M.I.T. differential analyzer.

For convenience in reference, the schematic diagram of the machine is repeated in Fig. 1. In the report mentioned, the utility of the middle part of the device was questioned. This arose from a misunderstanding of the basic assumptions underlying the design and was cleared up in a conference with Dr. Tappert. The writer's analysis was under the assumption that the mechanism was designed to find rates for linear forcing functions only (i.e., that higher order terms were small by comparison), and the analysis is still valid if this is true. However, in practice, it appears necessary to assume higher order forcing functions and the deflection mechanism is designed to give the correct steady state rate (except for the non-linearity of the sine gear) for an arbitrary quadratic forcing function. Actually the middle part (often referred to hereafter as the "x" part) of the device is certainly well worth while, as will be seen from some of our experimental curves.

If a linear mechanism has a transfer admittance $Y(j\omega)$ from input $e(t)$ to output $\dot{q}(t)$ then

$$j\omega Q(j\omega) = Y(j\omega)E(j\omega)$$

where E and Q are the transforms of e and q . It is easily seen from transform theory that if $e(t) = at + b$, a necessary and sufficient condition that $\dot{q}(t) \rightarrow a$ as $t \rightarrow \infty$ is that

$$Y(0) = D \left. \frac{dY}{d\omega} \right|_{\omega=0} = j$$

If this condition is satisfied the system may be called a first order rate finder — after the transient has died out, the output is the derivative of the input whenever latter is linear. Similarly if

$$Y(0) = 0 \quad Y'(0) = j \quad Y^{(k)}(0) = 0 \quad k = 2, 3, \dots, n$$

we have an nth order rate finder — in the steady state it finds the rate of an nth degree polynomial forcing function. In the deflection mechanism we have a second order rate finder

$$Y(j\omega) = \frac{1 + aj\omega}{1 + aj\omega + b(j\omega)^2 + c(j\omega)^3} j\omega$$

$$= j\omega + c_1\omega^3 + c_2\omega^4 + \dots$$

if we assume $\frac{1}{\sqrt{1 + \dot{q}^2}}$ nearly 1. A circuit for solving

$$\dot{q} = \sin^{-1} \dot{q}$$

under the same approximation, to the nth order is shown in Fig. 2. The admittance here is approximately

$$\frac{1 + a_1(j\omega) + a_2(j\omega)^2 + \dots + a_{n-1}(j\omega)^{n-1}}{1 + a_1(j\omega) + a_2(j\omega)^2 + \dots + a_{n+1}(j\omega)^{n+1}} j\omega$$

the values of the constants in the mechanism are

$$Y(j\omega) = \frac{1 + 4.63 j\omega}{1 + 4.63 j\omega + 5.73 (j\omega)^2 + 1.094 (j\omega)^3} j\omega$$

$$= \frac{(1 + 4.63 j\omega) j\omega}{(1 + .232 j\omega)(1 + 1.85 j\omega)(1 + 2.56 j\omega)}$$

In the previous report it was pointed out that due to a clutch and stop on the input to the sine gear values of \dot{q} were limited to two horizontal lines (see Fig. 6 in that report). There is also a clutch and stop on the displacement of the lower integrator. This effectively further limits solutions to a parallelogram as shown in Fig. 3. Actually the limitation is fictitious — the q shaft can turn an unlimited amount, but when this stop is in effect the stability point moves at such a speed as to be equivalent to q and \dot{q} moving along one side of the parallelogram. Thus if we keep the stable point stationary paths of representative solutions will be as indicated in Fig. 3.

The trial solutions taken on the differential analyzer may be classified as follows:

I. Solutions taken with the mechanism as designed.

A. Simple analytic forcing functions.

1. $e(t) = a$
2. $e(t) = at + b$
3. $e(t) = at^2 + bt + c$
4. $e(t) = at^3 + bt^2 + ct + d$

B. Response for 8 typical target courses, the target vector velocity constant.

C. The response to some error functions superposed on typical courses.

D. An attempt to get backlash oscillation.

II. Approximately the same program although less extensively with the middle part eliminated.

III. A few runs with typical courses using three different third order rate finders.

The constants of the target courses used were as follows (see Fig. 4):

Course I $S_g = 150 \text{ yds/sec} = 307 \text{ mi/hr}$

$V = 2,000 \text{ yds}$

$h_m = 1,000 \text{ yds}$

$\phi = 0^\circ$

Course II $S_g = 150 \text{ yds/sec}$

$V = 2,000 \text{ yds}$

$h_m = 500 \text{ yds}$

$\phi = 0$

Course III $S_g = 150 \text{ yds/sec}$

$V = 4,000 \text{ yds}$

$h_m = 1,000 \text{ yds}$

$\phi = 0$

Course IV $S_g = 150$
 $V = 2,000$
 $h_m = 2,000$
 $\phi = 0$

Course V $S_g = 150$
 $V_m = 4,000$
 $h_m = 4,000$
 $\phi = -14.96^\circ$
 $V = 4,000 - 40 t$

Course VI $S_g = 150$
 $V_m = 2,000$
 $h_m = 1,000$
 $\phi = -14.96^\circ$
 $V = 2,000 - 40 t$

Course VII $S_g = 96.6$
 $V_m = 3,000$
 $h_m = 1,000$
 $\phi = -50^\circ$
 $V = 3,000 - 115 t$

Course VIII $S_g = 150$
 $V = 4,000$
 $h_m = 500$
 $\phi = 0$

The distribution of these courses is indicated in Fig. 5, together with the approximate maximum range of the 3" A.A. gun (21 sec. fuse setting).

The actual input to the deflection mechanism is

$$\theta = \int_a^t \frac{S_g h_m}{h_o} \frac{t_p}{h_p} dt$$

but since it was desired to compare the actual output with the true deflection

$$\sin^{-1} \theta$$

the quantity θ was plotted against t and integrated to provide the input. To calculate θ the following method was found to be the simplest. We have

$$\theta = \frac{S_g h_m}{h_o} \frac{t_p}{h_p} \frac{1}{\sqrt{1 + \left(\frac{S_g}{h_m} t\right)^2}} \frac{t_p}{h_p}$$

A computation schedule was set up based on this formula, working backwards from the time of burst $t + t_p$ to the present time

I (assumed)	II	III	
$t + t_p$	h_p	V_p	
	$= h_m \sqrt{1 + [S_g (t+t_p)]^2}$	$= V_m [1 - (t+t_p) S_g \tan \phi]$	
IV	V	VI	VII
t_p	t	$\sqrt{1 + \left(\frac{S_g}{h_m} t\right)^2}$	θ
from ballistic curves	$= I - IV$		

The ballistic data used in getting t_p (IV) was read from the chart Fig. 24 (opposite p. 59), Coast Artillery Field Manual, FM 4-110. The value of t_p was merely read off corresponding to the computed values of r_p and h_p .

If we assume as an approximation that the shell velocity is constant, k yds/sec (i.e., that the equi-time of flight curves in the chart are circles) so that with V constant

$$k^2 t_p^2 = h_p^2 + v^2$$

$$h_p = h_m + S_g(t + t_p)$$

$$\dot{e} = \frac{S_g t_p}{h_p \sqrt{h_m^2 + S_g^2 t_p^2}}$$

we can eliminate t_p and h_p from the system to obtain the following equation between e_0 and t :

$$\begin{aligned} & \dot{e}^2 [k^2 (h_m + S_g t)^2 (h_m^2 + S_g^2 t^2) - (h_m^2 + S_g^2 t^2) v^2 S_g^2] \\ & + \delta [2 \sqrt{S_g^2 h_m} \sqrt{h_m^2 + S_g^2 t^2}] - [\sqrt{S_g^2 h_m^2 + S_g^2 h_m^2 (h_m + t S_g)^2}]^2 = 0 \end{aligned}$$

Evidently the same curve $e_0(t)$ is obtained if h_m and S_g are both multiplied by the same constant.

The differential analyzer set-up used is shown in Fig. 6. An attempt was made to generate the sine function with two integrators solving

$$\frac{d^2 \dot{q}}{dt^2} = -\dot{q}$$

but this was found impractical because of the large integrator loading necessary, and an input table was used instead. Even in this case it was necessary to use a very large scale factor on the independent variable shaft due to the small integrating factors ($1/32$) of the differential analyzer as compared to the ball type (about 1 under comparable conditions). This resulted in solutions which represented, actually, 30 seconds requiring 30 minutes of machine time.

The equations of the deflection mechanism are

$$\begin{aligned} \ddot{x} + .54 \dot{x} &= .54 \dot{e} \\ \sqrt{\frac{q}{1-\dot{q}^2}} + 4.700 \dot{q} + 1.692 q &= 1.692 e + 4.700 x \end{aligned}$$

It was necessary to approximate the coefficients with available gear ratios on the differential analyzer. Fortunately some very close approximations were found. The equations actually set on the machine were

$$\dot{x} + .54 x = .54 \dot{e}$$

$$\frac{\ddot{q}}{1-\dot{q}^2} + 4.706 \dot{q} + 1.694 q = 1.694 e + 4.706 x$$

The error is of the same order as the expected machine error.

Except for runs in group ID the machine was made as "tight" as possible, the backlash being corrected by frontlash units. Due to the large scale factors used and the high inherent precision of the integrators used in the differential analyzer, the runs may be expected to be more accurate than the actual deflection mechanism.

Solutions were taken in the form of both curves and counter readings. The curves given here were reproduced by pantograph to ordinary graph paper size. Curves not directly drawn by the machine and numerical values quoted are taken from the counter printings, which give an additional decimal place not readable from the curves.

Discussion of Runs

Most of the curves are given with \dot{q} as dependent variable. To estimate the error in yards for a given error in \dot{q} from \dot{e} , the chart of Fig. 6A may be used. This is computed from the approximate formula

$$\begin{aligned} r \cos \epsilon \Delta \dot{q} \\ = r \frac{\cos \epsilon}{\sqrt{1-\dot{q}^2}} \Delta \dot{q} = r A(\epsilon, \dot{q}) \Delta \dot{q} \end{aligned}$$

For rough comparisons the coefficient A may be taken as 1, the error then being the \dot{q} error multiplied by the predicted range.

The first set of runs taken were with a sudden impulse $e = k1$ with the system at rest, both with and without the middle part of the mechanism. Runs were taken with

$$k = 0.1, 0.2, 0.4, 1.0, 2.0$$

Typical curves are shown in Figs. 7 and 8. The results are very close to computed curves on the assumption that $1/\sqrt{1-\dot{q}^2} = 1$ when $k \leq .4$, but above this the non-linearity becomes appreciable. In the worst cases the transient disappeared to within machine errors in 25 seconds, and for most cases within 8 to 12 seconds. The action with the middle part out was

considerably more rapid than with it in, the transient being 6 times as great, as had been predicted, this being a special case of a linear forcing function. Fig. 9 is a plot of the time required for the transient in \dot{q} to reduce to 2/10 of its maximum value. For values of k greater than about .35 the curves cross the axis once with the middle part in. The curves with it out are all identical with $k \geq 2$, due to the action of the slip clutch on one integrator.

Next a series of runs were taken

$$e = kt_1(t)$$

starting from rest, with

$$\sin^{-1} k = \text{steady state } \delta = 15^\circ, 30^\circ, 45^\circ, 60^\circ, 75^\circ, 80.6^\circ$$

the last being the limit of the sine gear, the maximum possible deflection. These runs are shown in Figs. 10 and 11. The transient died out in all cases within 20 seconds except with x in for $\delta \geq 75^\circ$ in which cases 30 seconds or more was required, due to the action of the slip clutch. These long transients, however, would probably not be troublesome since such large deflections would only occur in practice with the plane almost directly overhead. For the smaller values the response is about equally rapid with x in or out.

Quadratic Forcing Functions

The runs with a quadratic forcing function

$$e = at^2$$

were the first to show the superiority of the mechanism with x in. Runs were taken with

$$a = .01, .02, .03, .04, .10$$

With a quadratic rate finder the solution \dot{q} should approach 2 at, and with x in this was very nearly true, the discrepancy being due to the sine gear. Some solutions are shown in Figs. 12, 13, and 14. The errors increase with a and with \dot{q} . The maximum slope found in any of the \dot{q} courses plotted is about equivalent to an a of .05 so that the large errors due to the sine gear with $a = .10$ need not cause great concern.

Cubic Forcing Functions

For cubic forcing functions the following were used

$$e_1 = -.04 t^3 + .1 t^2$$

$$e_2 = -.001 t^3 + .05 t^2$$

$$e_3 = -.0002 t^3 + .02 t^2$$

These were chosen as having second order tangency at $t = 0$ so that the transient is small. The results are shown in Figs. 15 and 16. The response with e_2 and especially e_3 are very close to the calculated values on assuming the equation linear. The error in e_3 is somewhat greater as in the quadratic case with higher acceleration.

Effect of Backlash

A number of runs were made to determine the effect of backlash using several different forcing functions. In order to increase the amount of backlash, frontlash units were inserted at several critical points in the backwards direction. The results of these runs were, however, completely negative, for no oscillation of any sort was discovered. The system was given "shocks" by sudden turning of the \dot{e} shaft and other methods, but the solutions were completely stable. The only results were small consistent errors, of the order of magnitude of the backlash. It is possible that due to the large scale factors used in the set up, even the artificially introduced backlash was not sufficient to cause the oscillation effect.

Response for Typical Courses

The response for the 8 courses described above are shown in Figs. 17 to 24. It may be noted that even on the flat courses (e.g., IV) the operation is poor without x . On the flat courses the response is satisfactory with x , the error being less than 20 yards except sometimes at the hump in \dot{e} . However for the steeper courses errors of 50 or more yards are common after the start of the peak which do not disappear until nearly the end of the course. The action is particularly bad coming down the hump. Fig. 25 is a plot of the error in yards with course VIII, x in.

Response to Error Functions

In Figs. 26 - 28 are shown the responses to some random error functions of various kinds superimposed on courses I and II. The operation in damping out the error is considerably better with x out. However it seems from a consideration of the size of the errors introduced and the responses found that the system, even with x in, damps the errors more than necessary. That is, it might be preferable to increase the speed of response so as to reduce the transient errors in the solutions.

Figs. 29 and 30 show the responses when we suddenly start tracking a target in courses I or II with the machine previously at rest, with the target at several points along the course.

Tests with Different Equations

Three runs were made on course VIII, the most difficult one of the group, using three different cubic rate finding equations. The equations used were (assuming linearity) critically damped, with the transfer admittances:

$$(1) \quad \dot{q} = \frac{1 + 8(j\omega) + 24(j\omega)^2}{[1 + 2(j\omega)]^4} j\omega e$$

$$(2) \quad \dot{q} = \frac{1 + 4(j\omega) + 6(j\omega)^2}{[1 + (j\omega)]^4} j\omega e$$

$$(3) \quad \dot{q} = \frac{1 + 2(j\omega) + \frac{3}{2}(j\omega)^2}{[1 + \frac{j\omega}{2}]^2} j\omega e$$

The results of these runs are shown in Figs. 31, 32, and 33 and should be compared with Fig. 24. Of course, this gain is accompanied with a loss in error function damping. With the roots equal to 2 the system had a slight tendency to be unstable on the flat part of the course. This however appeared to be due to the "human backlash" in the operator on the sine table and would probably not be present with a sine gear.

It is easily seen that an increase in the values of the characteristic roots of the equation demands a proportional increase in the power requirements of the integrators. It may be that this will be a design limit in the case of mechanical systems. No difficulty would be experienced here however with electrical integrators.

The main conclusions of this work are as follows:

1. The middle part of the machine is definitely worth while. Although it increases response for accidental following errors, the gain in behavior for actual courses more than offsets this disadvantage.

2. The system behaves nearly enough like the linear system

$$1.094 \ddot{q} + 5.73 \dot{q} + 4.63 q + \dot{q} = 4.63 \ddot{e} + 4.63 \dot{e}$$

that this may be used to calculate its response to within a few per cent, providing $\dot{q} < .6$. As this corresponds to a deflection of 37° , the approximation is sufficient for most cases.

3. For targets whose elevation at their nearest point is greater than about 50° fairly large errors occur due to substantial cubic and higher degree terms in e . This indicates that it might be worth while to use a higher order rate finder. Tests made with a cubic rate finder showed greatly improved results.

4. If the additional cost of another integrator and adder required for cubic rate finding is too great to be justified it appears that the system could be improved by reducing the time constants, for if sufficient power is available from the integrators, the only disadvantage would be increased response to random error functions and our results indicate that they are now damped out more than necessary.

5. There is some indication that better results would be obtained by making the three time constants equal, or more nearly equal than they are now, although this is not certain.

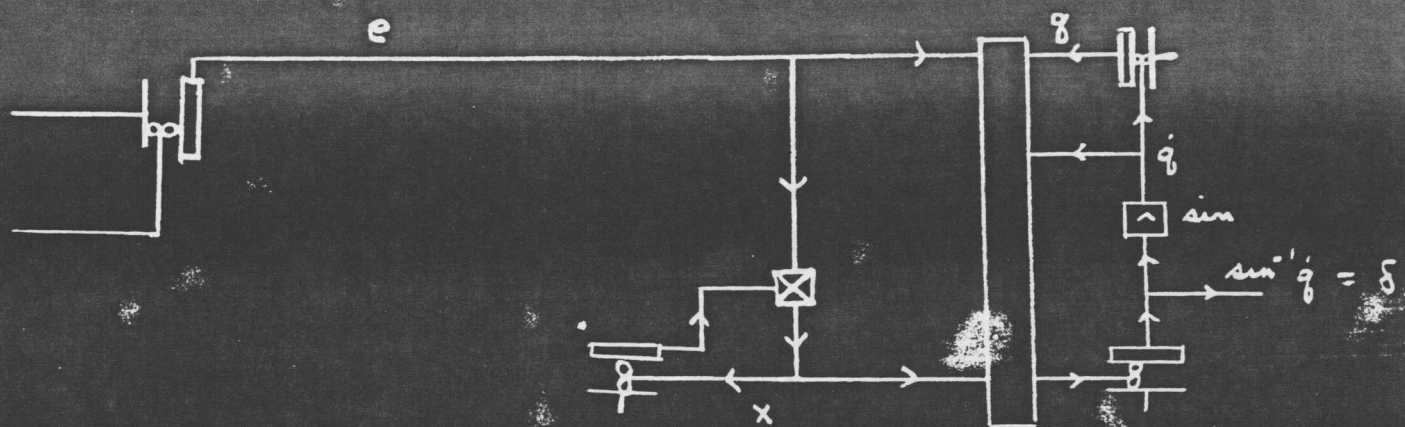


FIG. 1.

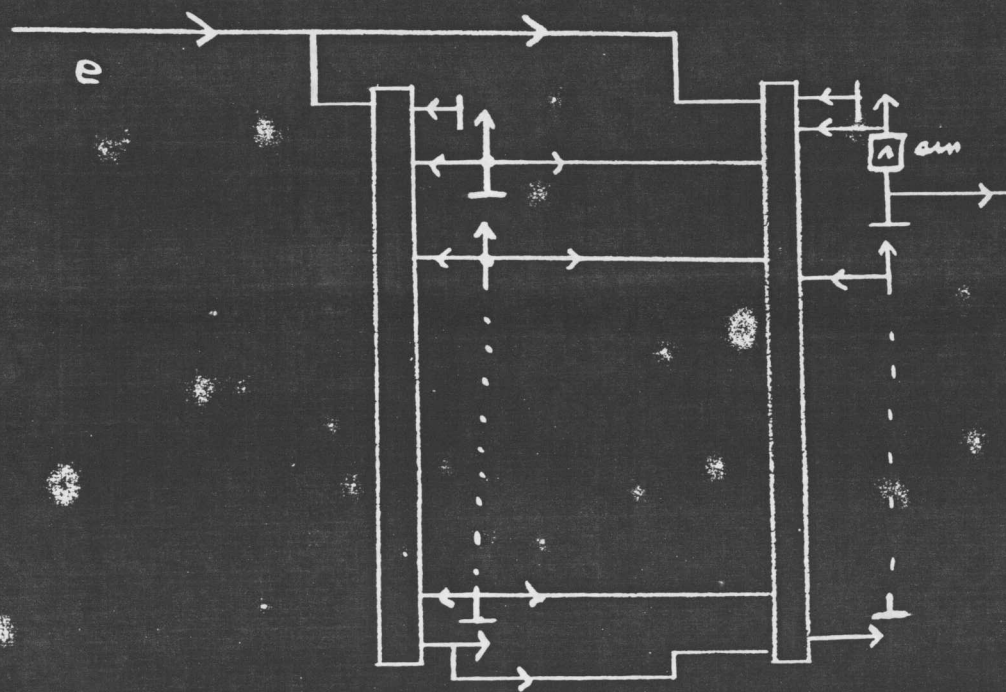


FIG. 2.

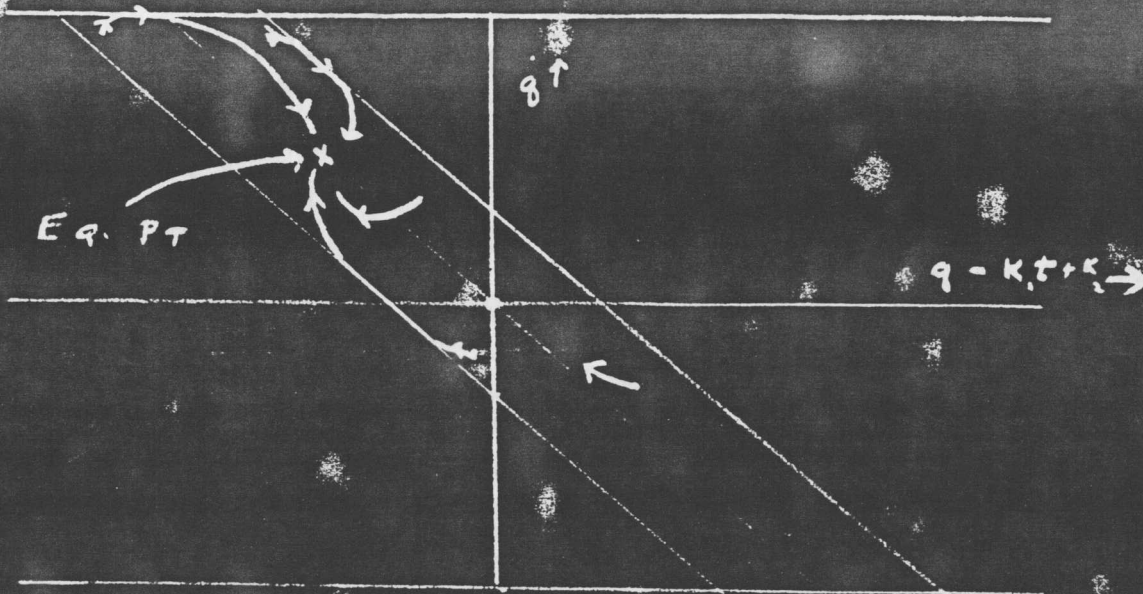


FIG. 3.

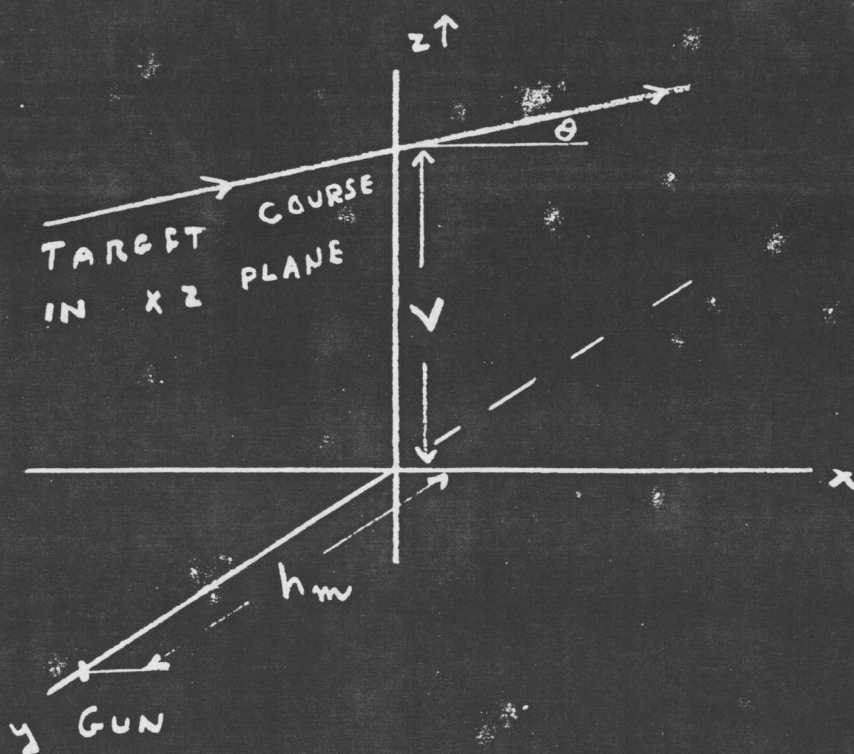
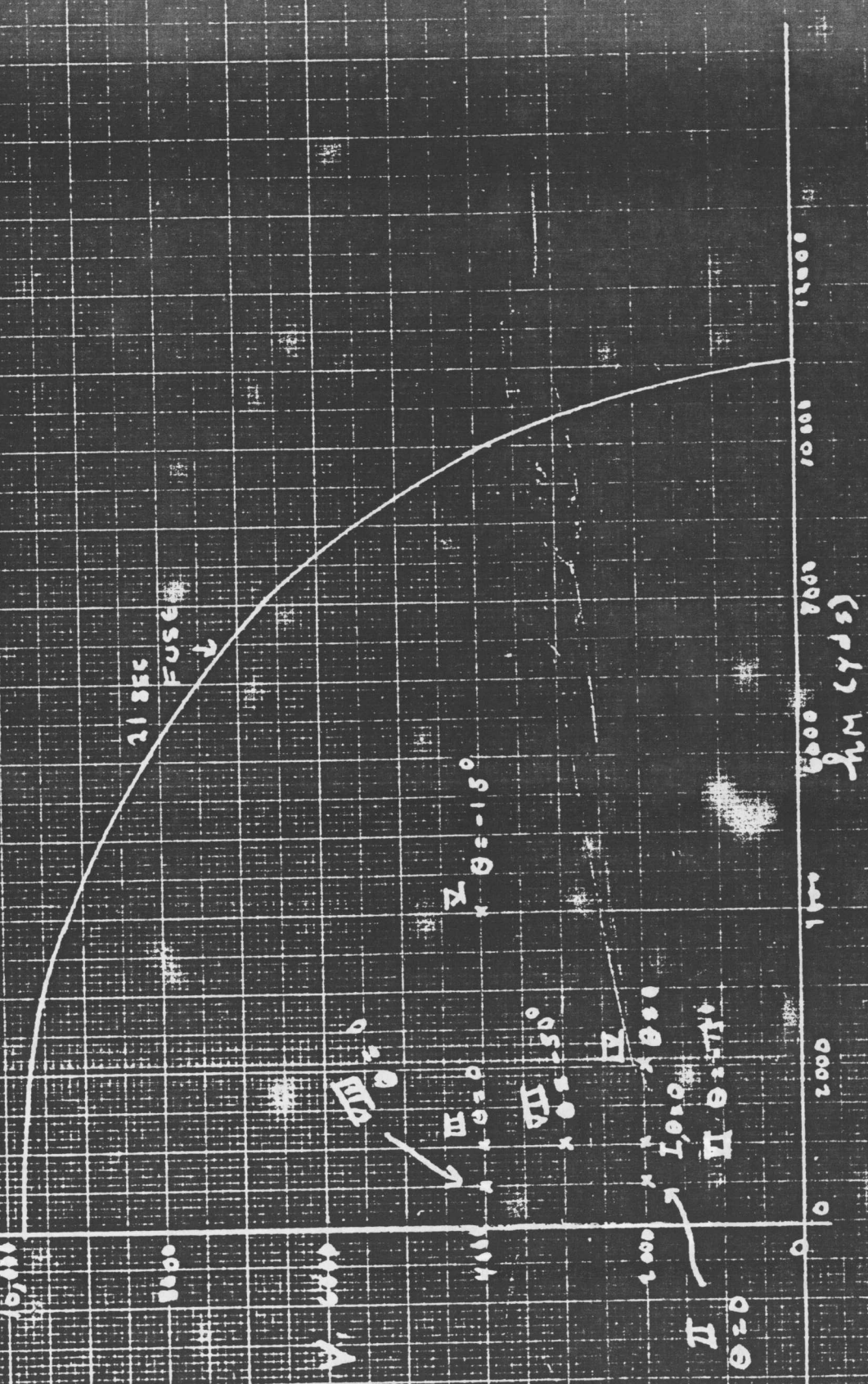
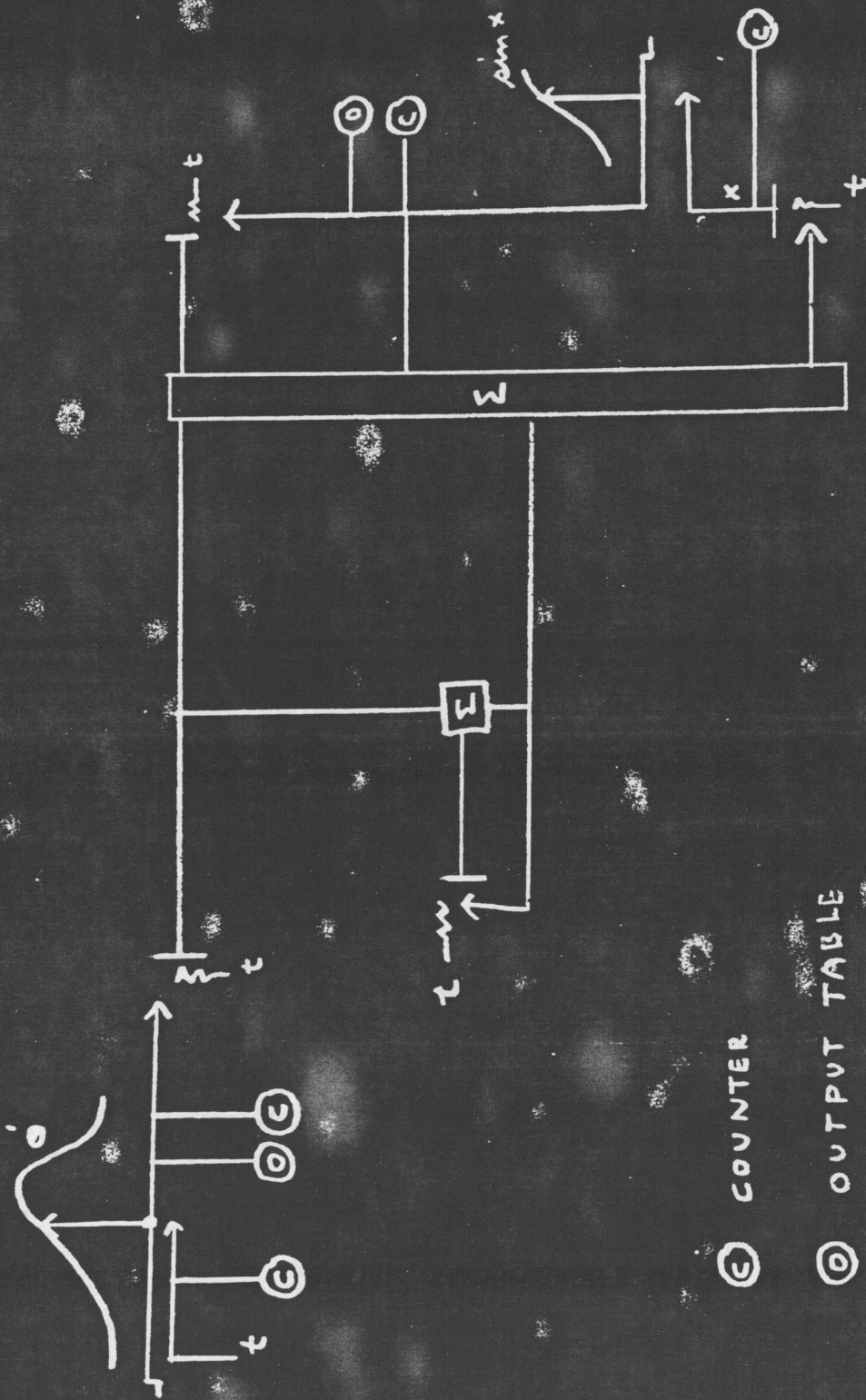


FIG. 4

FIG. 5

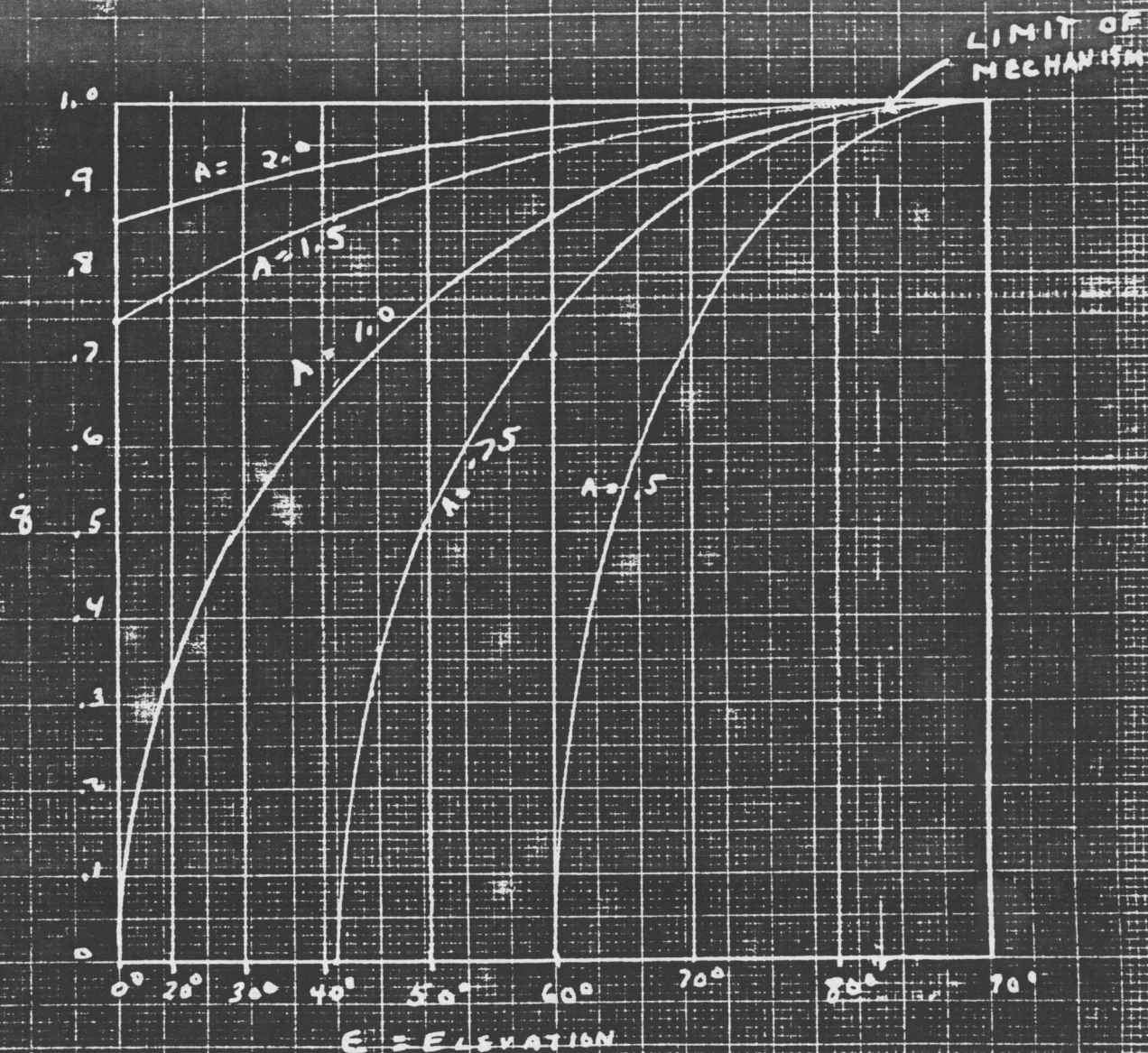




⊙ COUNTER

⊙ OUTPUT TABLE

FIG. 6.



ERROR (YDS) = RANGE (YDS) *
(ERROR IN g)

FIG. 6A

Fig 7

$\rho = C I(t)$

$C =$

← CS 2.0 or more

CS 1.0

CS 1.0

CS 1.0

CS 1.0

5.0

2.5

0

17.5

10.0

7.5

50

25

25

0

Form No. 99D-20 Squares to Inch
AMERICAN PAD & PAPER CO. HOLYOK, MASS.

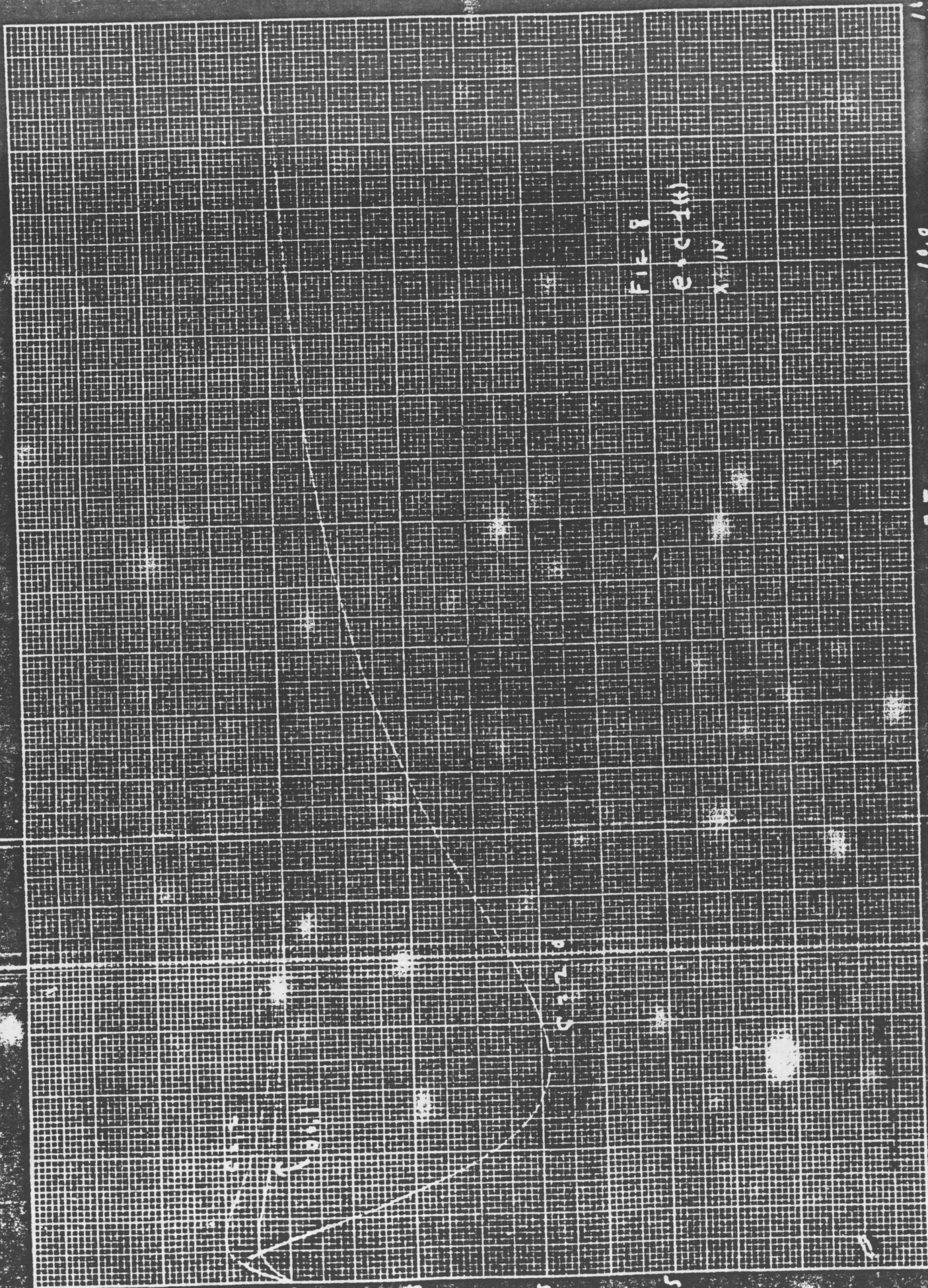


FIG 8

e-c-10

X 1/2

632

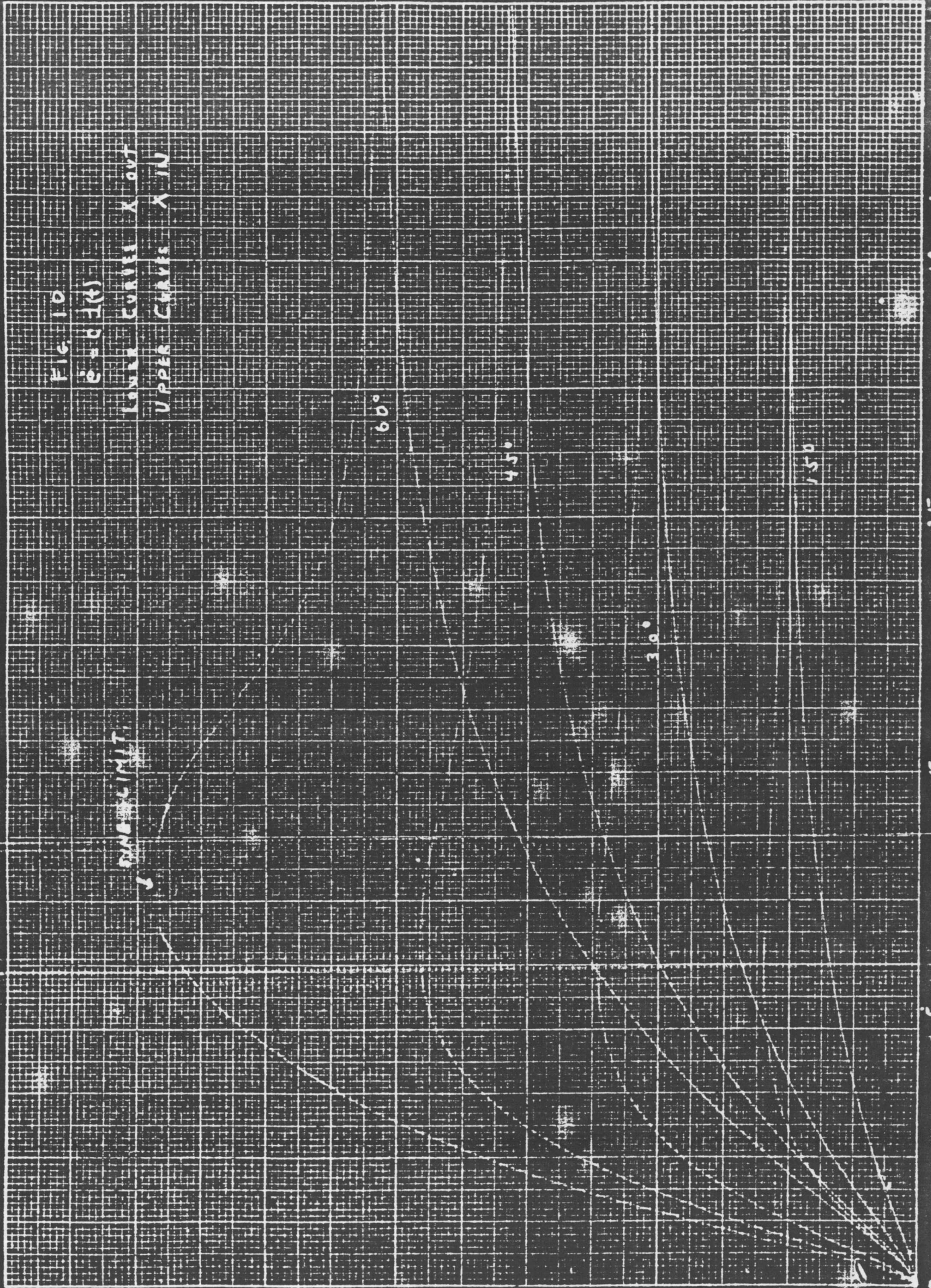
MIDDLE PART IN

MIDDLE PART OUT

FIG. 9

TIME TO REACH
 $\frac{1}{10}$ OF MAXIMUM,
 $t = K \ln(10)$

3.2 3.0 2.8 2.6 2.4 2.2 2.0 1.8 1.6 1.4 1.2 1.0 0.8 0.6 0.4 0.2 0



τ (LAST PART OF 750 X IN CURVE)



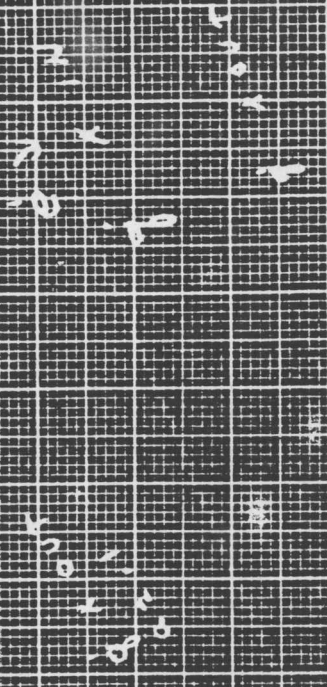
Fig 3

$Q = 0.27$

$Q = 0.27$

STARTING IN APPROXIMATE
STEADY STATE

NEW LIMIT



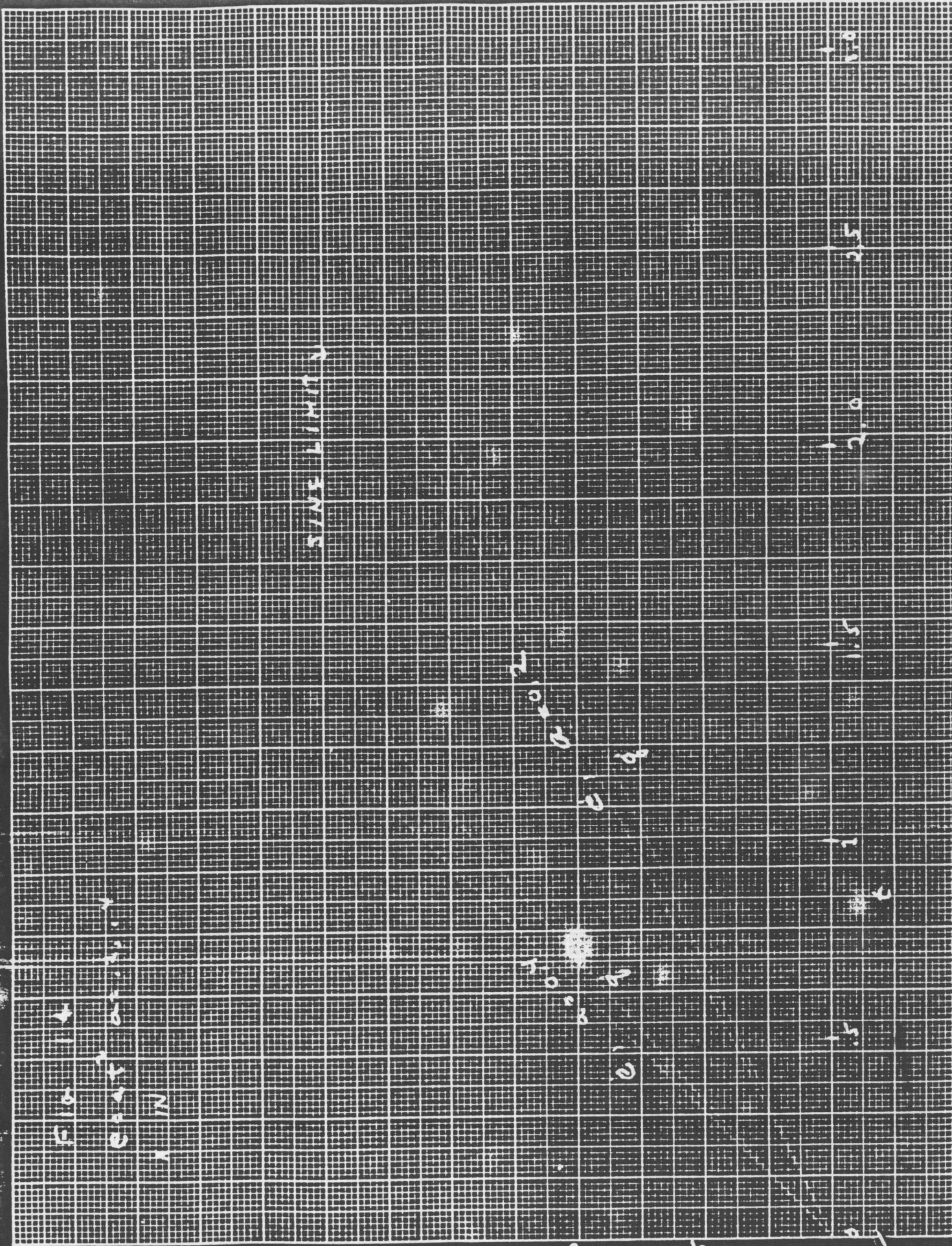
1.0

0.75

0.50

0.25

0.0



5/11/11

1.0

1.5

2.0

1.5

1.0

1.5

0.5

1.0

1.75

2.0

1.5

2.0

1.5

0.5

Fig 1/5

(UNDER AXIS)

Fig 2

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

Fig 1/5

1/2
COURSE T

1/2
COURSE T

1/2

1/2

1/2
COURSE T

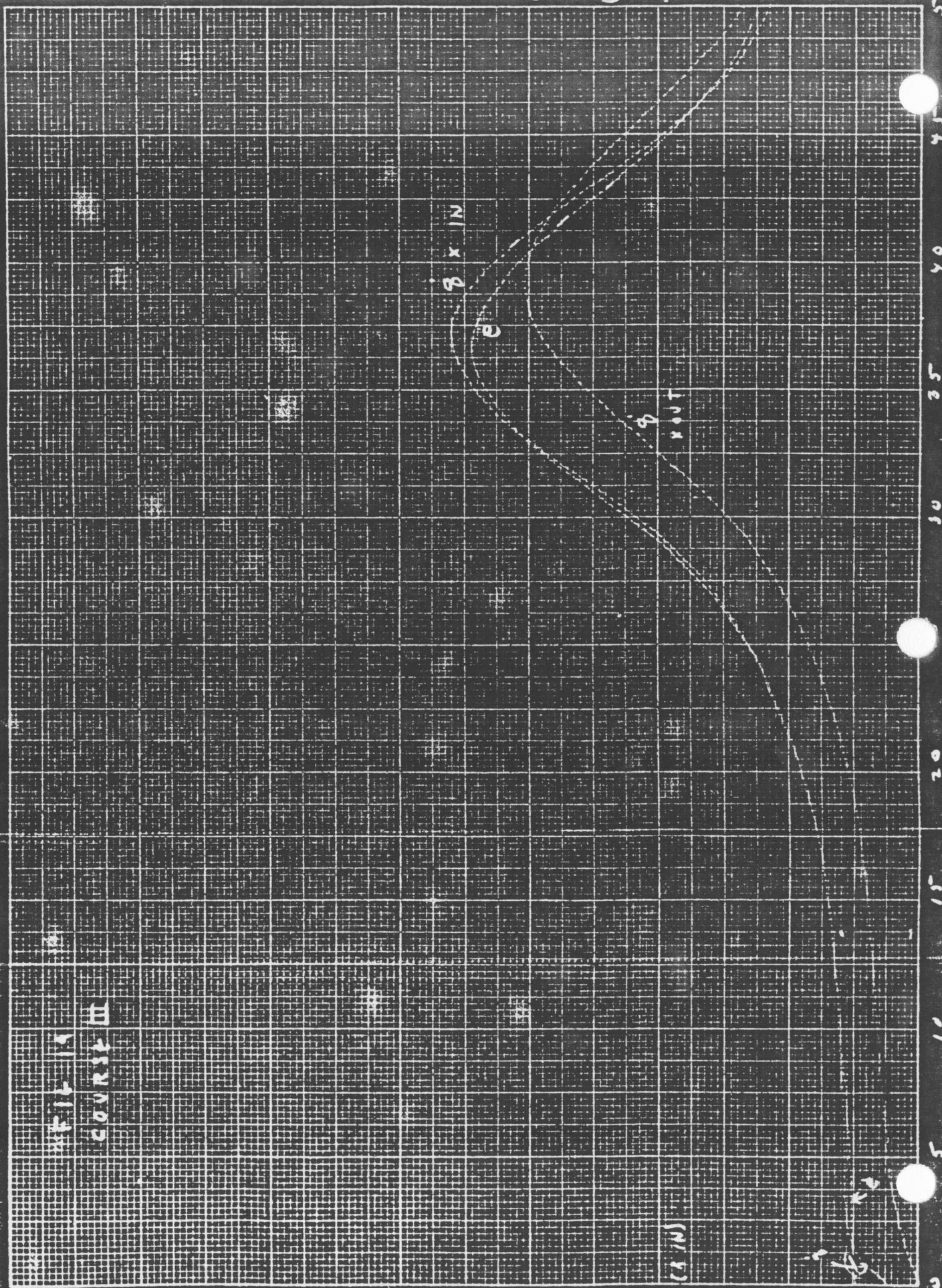


FIG. 10
COURSE IV

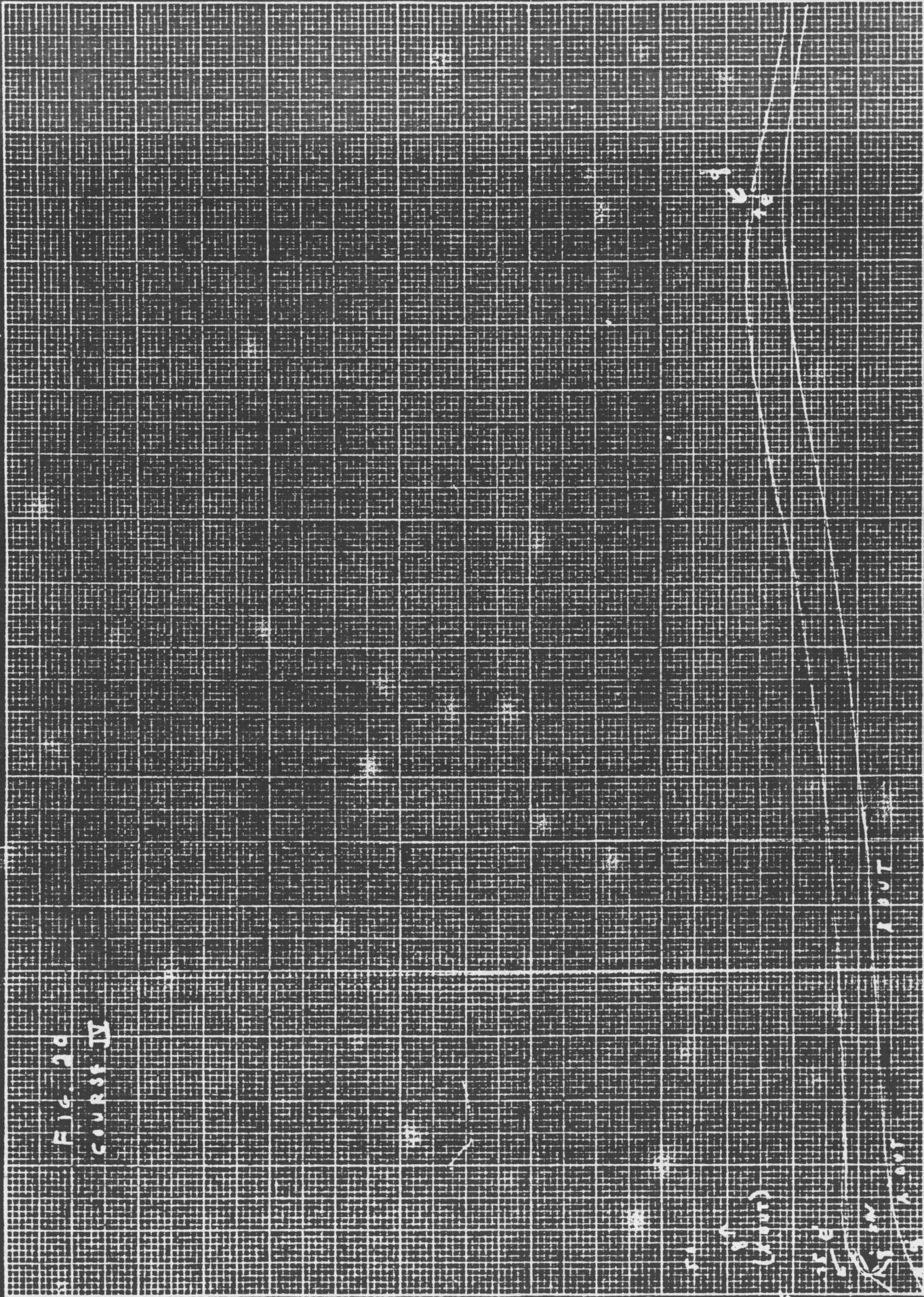


FIG 21

GEORGE V

21
6
10
15
20
25
30
35
40
45
50

1005

0
5
10
15
20
25
30
35
40
45
50

-30 -25 -20 -15 -10 -5 0

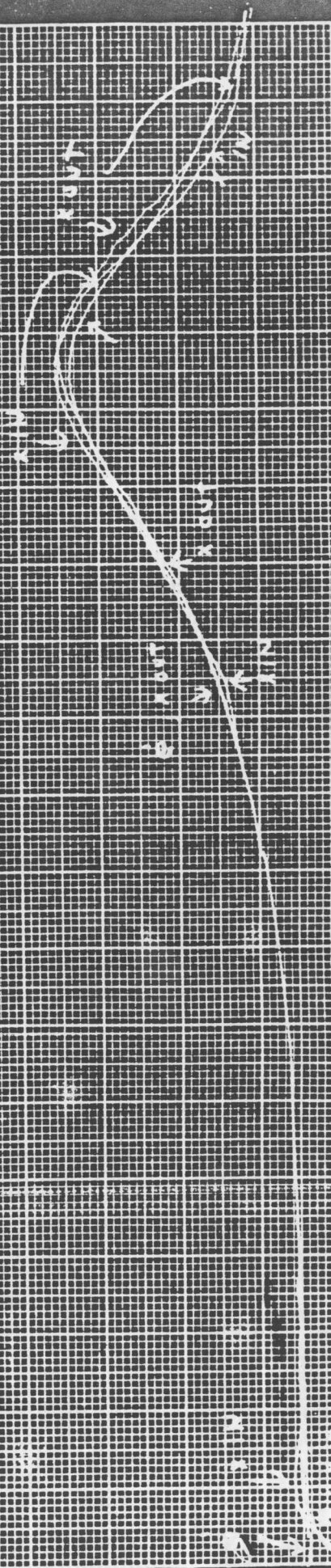
21
6
10
15
20
25
30
35
40
45
50

F 14 322

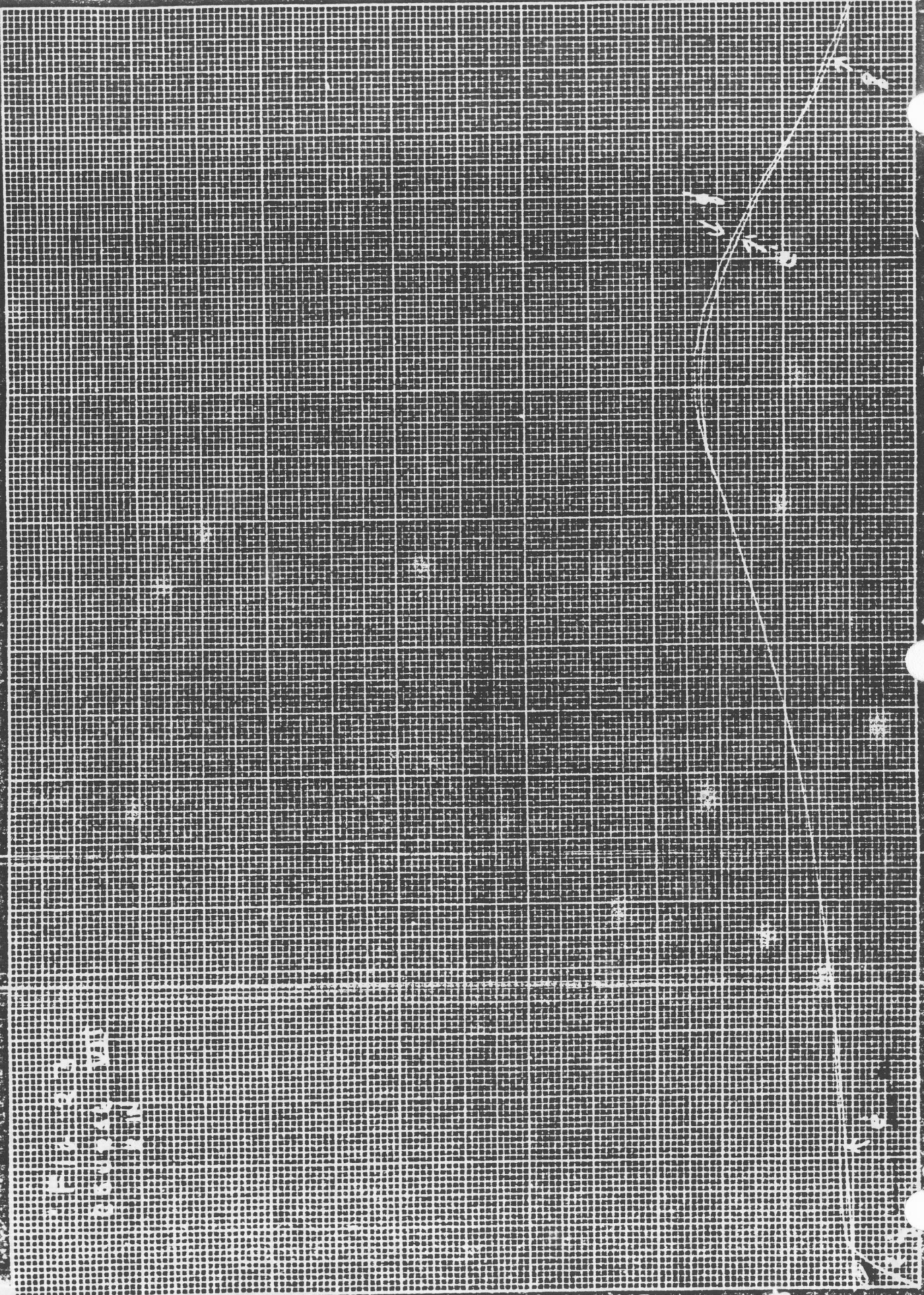
CO. 0. 55 111

130

15

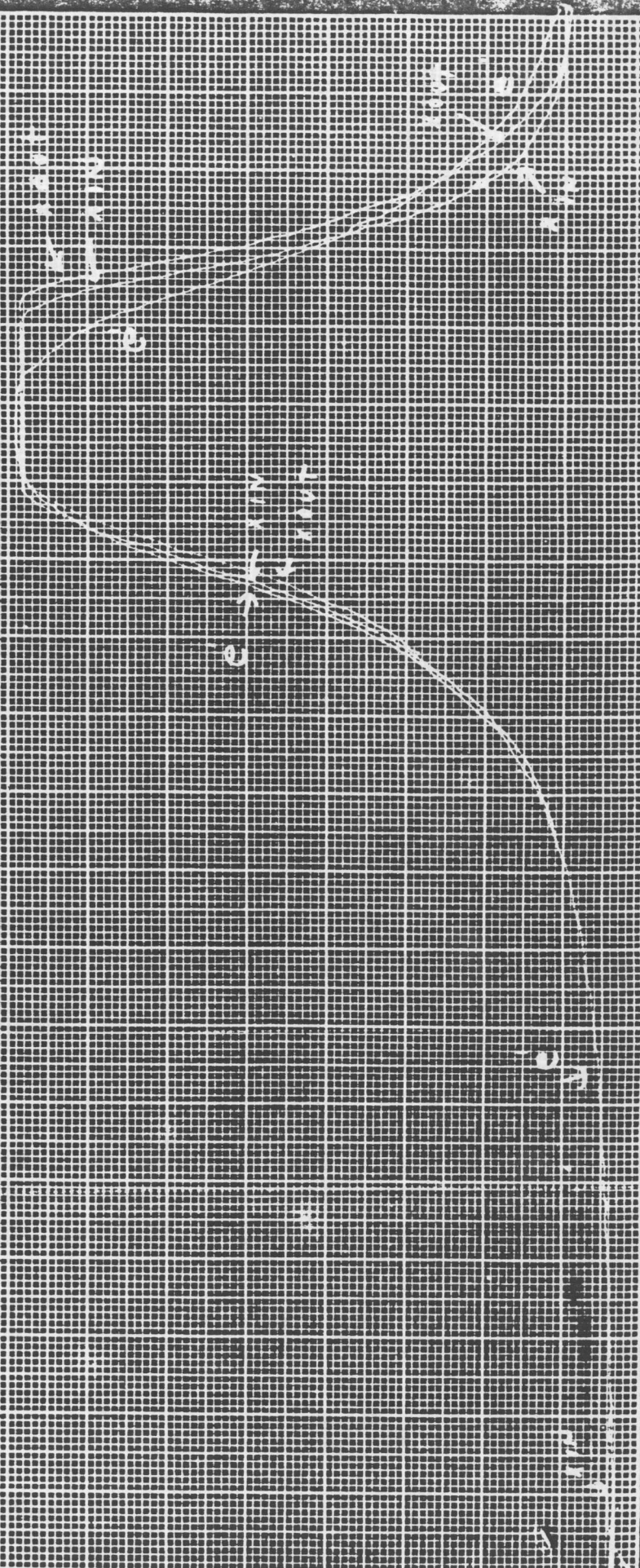


Form No. 909-20 Square to Inch
AMERICAN CAM PAD & PAPER CO. HOLYOKE, MASS.



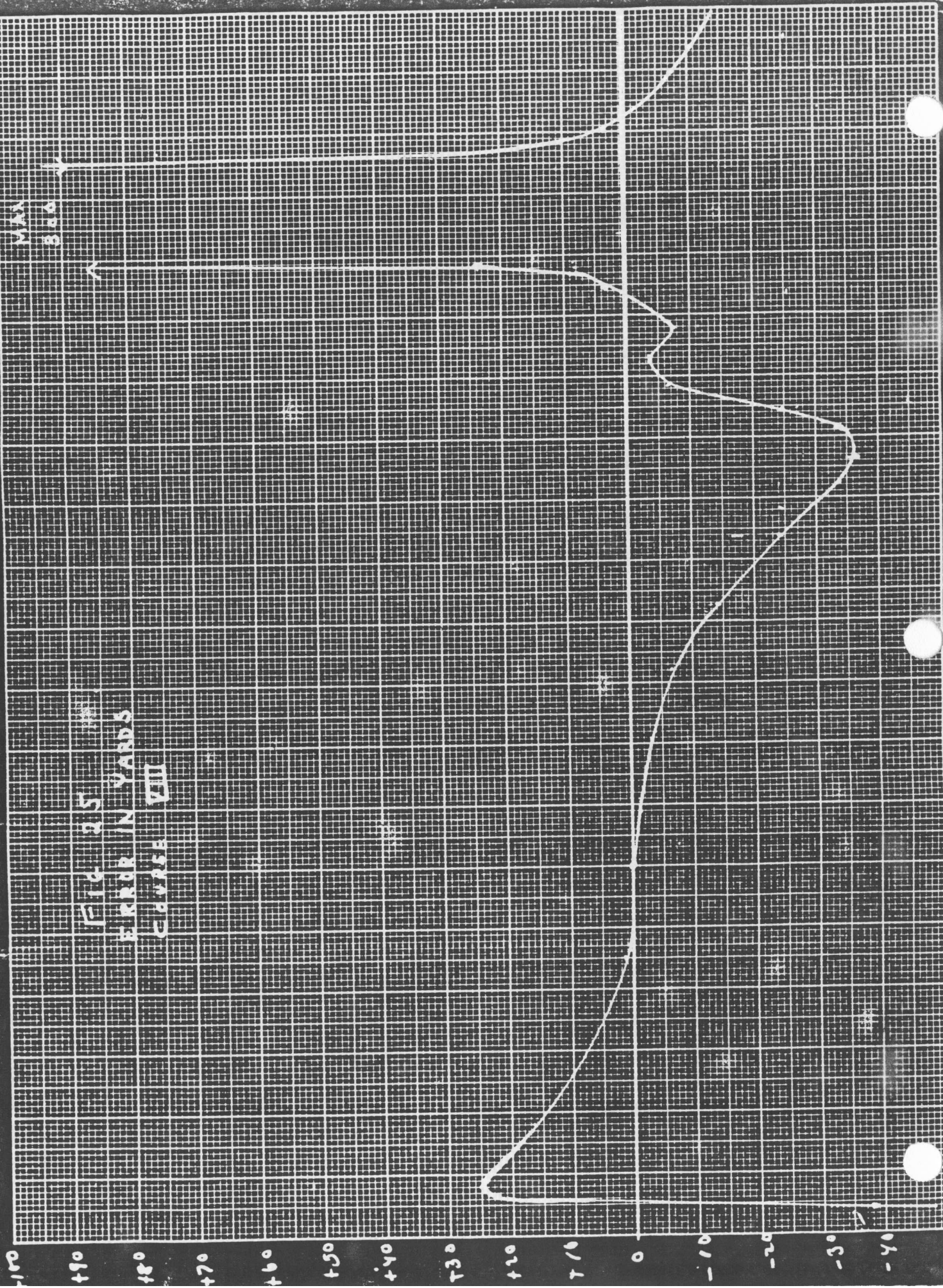
KUPFER & ESSER CO., N. Y. NO. 319-120
18 X 10 to the half inch.
MADE IN U. S. A.

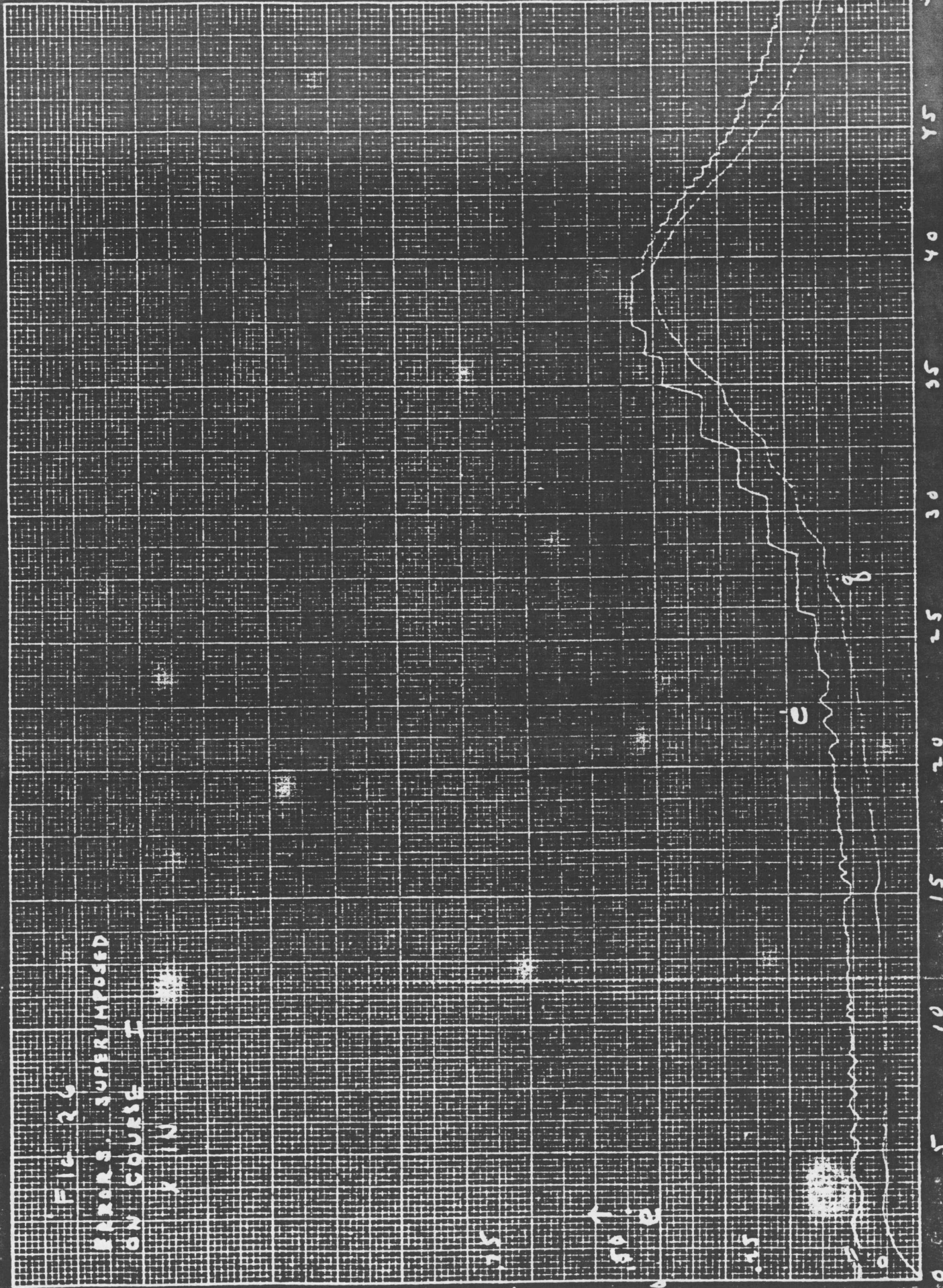
PLATE NO. 1
COLUMBIA



KEUFFEL & ESSER CO., N. Y. NO. 389-150
10 x 10 to the half inch.
MADE IN U.S.A.

1516.25
ERRORS IN YARDS
GRADE VIII





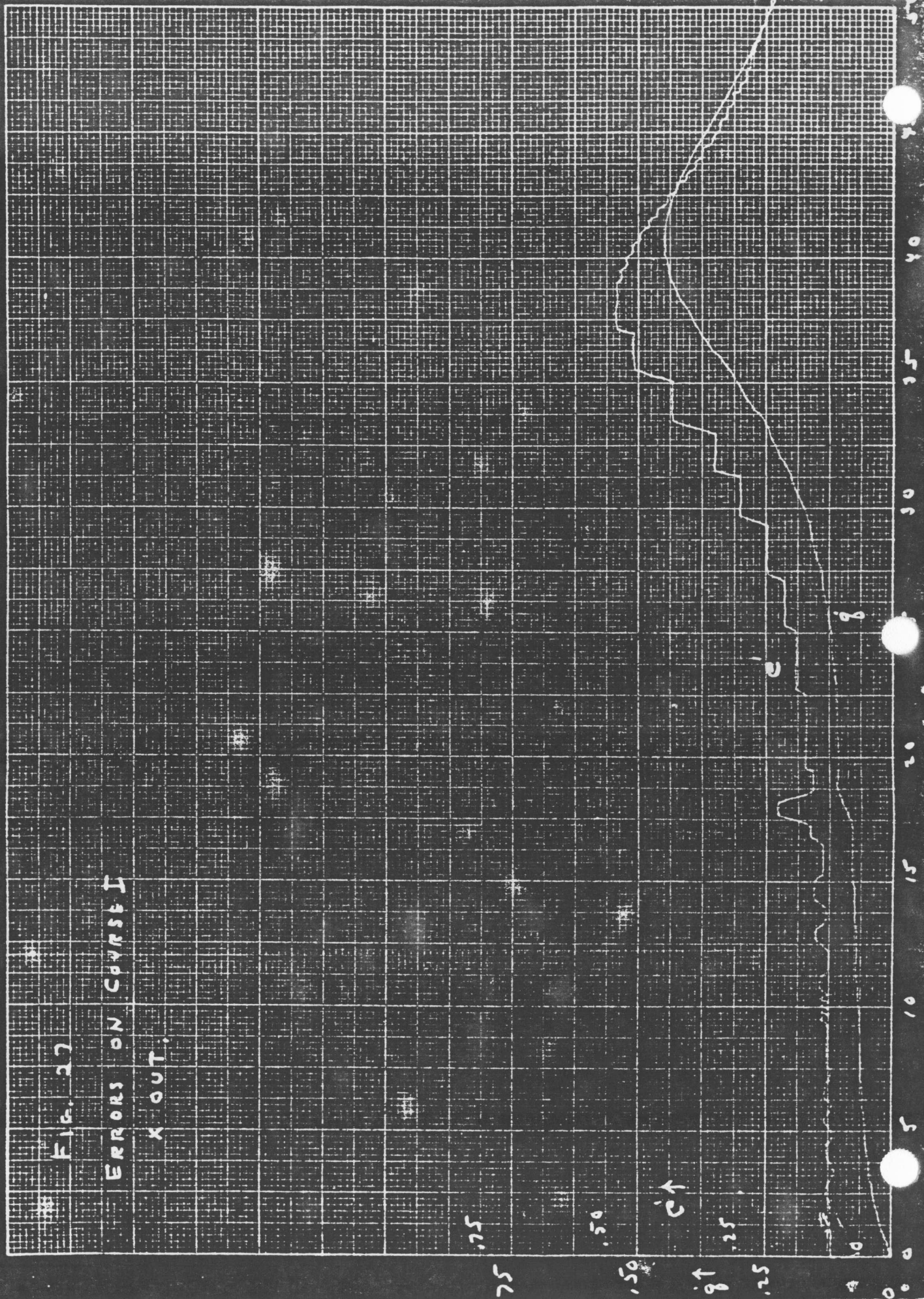
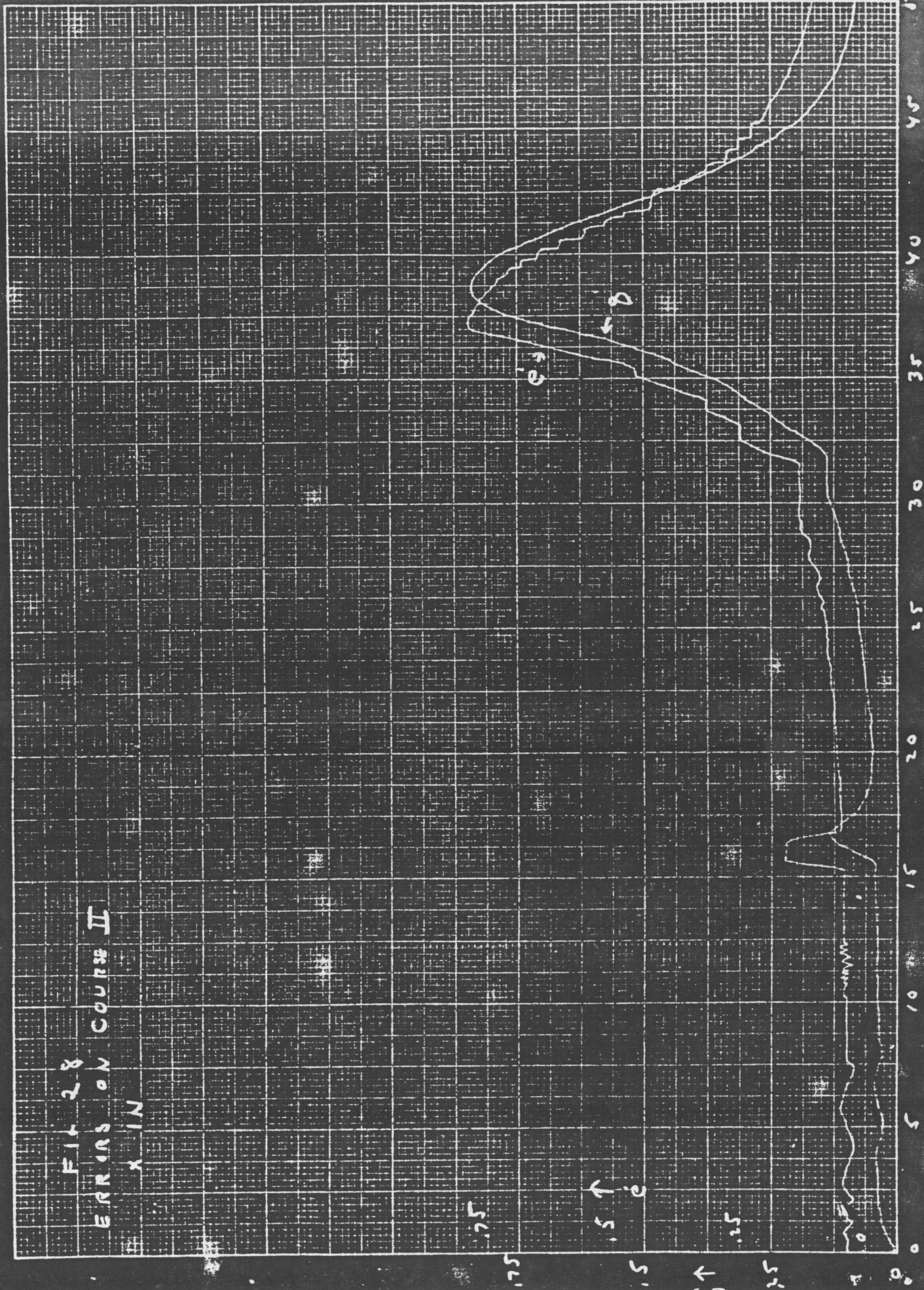


FIG. 28
ERRORS ON COURSE II
X IN



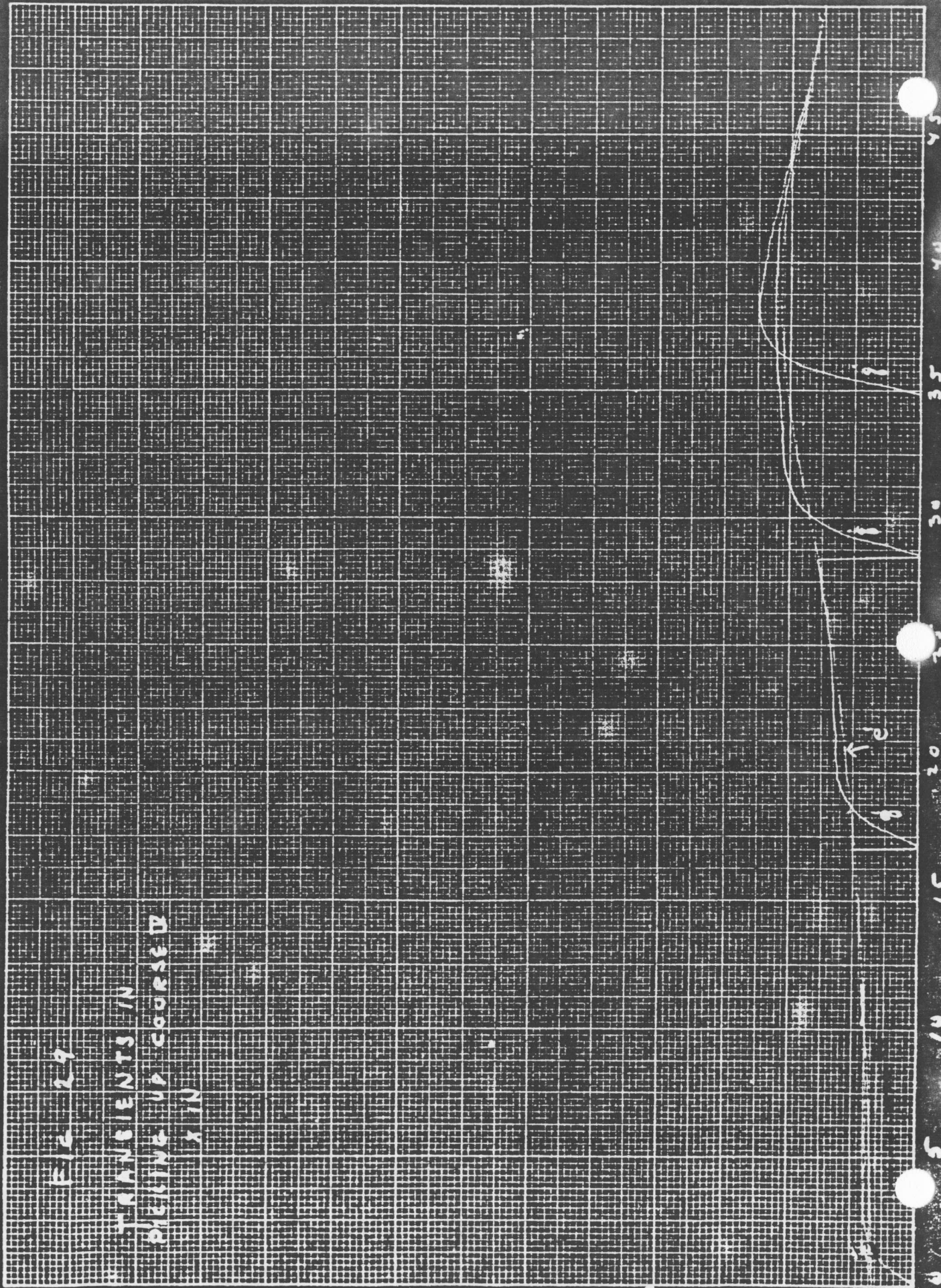


Fig 30

TRANSIENTS IN

PICKING UP COAST III

KIN.

4.9

175

170

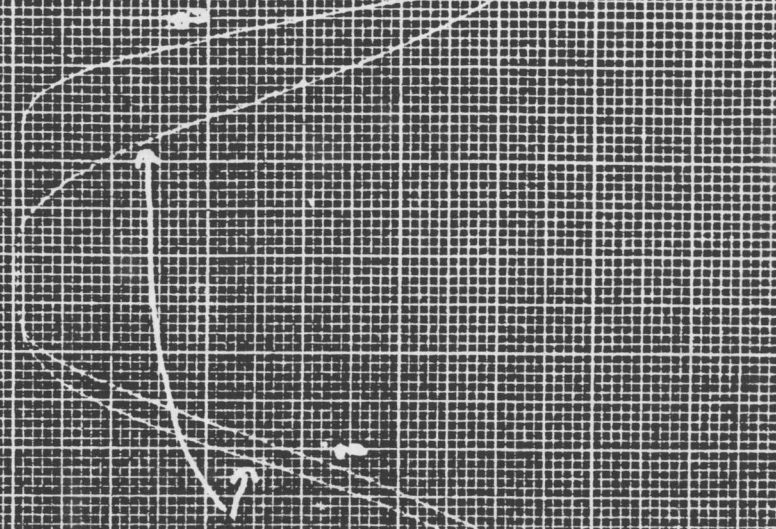
165

160

15 20 25 30 35 40 45 50

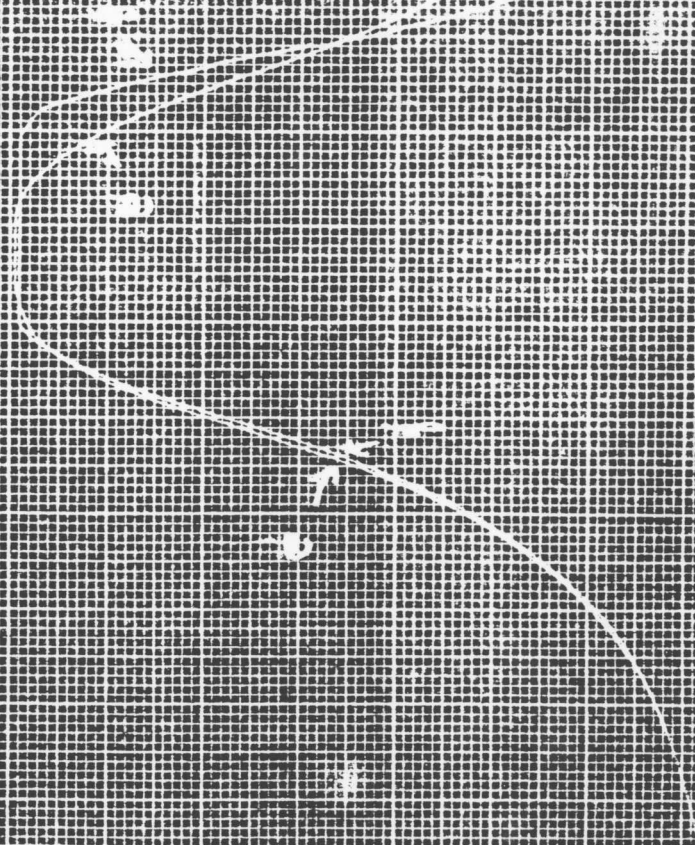
FIG. 21
CABLE ROUTE FINDER
HOLYOKE, MASS.
CABLE ROUTE FINDER

FIG. 21



Form No. 990-20 Squares 15 Inch
AMERICAN PAD & PAPER CO., HOLYoke, MASS.

FILE 12
CUBAN REVOLUTION
MARTIN LUTHER
KING, JR.
COLORED PEOPLE



Criteria for Consistency and Uniqueness in Relay Circuits

[12]

September 8, 1941

In a system of linear algebraic equations, there are three possible types of degeneracy, namely inconsistency (no possible solution), ambiguity (solutions not uniquely determined) and redundancy (more equations than necessary). Necessary and sufficient conditions are known for these types of degeneracy in terms of the ranks of ^{the} coefficient and augmented matrices. Somewhat similar effects can occur in the Boolean equations characterizing relay circuits, giving rise respectively to shattering, ambiguity of relay position for certain values of the independent variables, and redundancy of relays or contacts. In this note partial criteria will be established for these conditions in terms of a circuit discriminant φ .

Consider a relay circuit containing n relays R_1, R_2, \dots, R_n . Make and break contacts on R_i are designated x_i and \bar{x}_i , and we suppose that there are m independent variables a_1, a_2, \dots, a_m , which do not depend on the relay positions. Such a circuit is equivalent to the circuit of Fig. 1 in which

$$R_1(a_1, a_2, \dots, a_m; x_1, x_2, \dots, x_m)$$

is the Boolean function which is zero when the switches

and contacts a_1, \dots, x_n are in such positions that the voltage across R_1 in the original circuit is sufficient to operate it and one otherwise. The function

$$\varphi = \varphi(a_1, \dots, a_n, x_1, \dots, x_n) = \sum_{i=1}^n x_i \otimes R_i(a_{11}, \dots, a_n, x_1, \dots, x_n)$$

will be called the circuit discriminant. We also define the following items. A steady state in a relay circuit corresponding to a given set of values of the independent variables A_1, A_2, \dots, A_n is a set of positions P_1, P_2, \dots, P_n of the relays such that if the independent variables are given the values A_1, \dots, A_n and the relays held in the position P_1, \dots, P_n long enough for the steady state fluxes in the coils to build up, the relays will remain in the same positions indefinitely.

A completely oscillatory state of a relay circuit is a set of values A_1, A_2, \dots, A_n of the independent variables, such that no matter what the initial positions of the relays, or how long they are held in that position, when they are released at least one makes an infinite number of oscillations, i.e. chatters. In addition to these obviously exclusive possibilities a circuit may be "partially" oscillatory for certain

values of the independent variables- with some initial conditions the circuit chatters and with others relapses into a steady state. An example is shown in Figure 8 where with the initial condition

$$R_1 = 0 \quad (\text{operated})$$

the circuit chatters while with

$$R_1 = 1$$

the circuit relapses into the steady state $R_1 = 1, R_2 = 1$

Theorem I - For $A_1, \dots, A_n; P_1, \dots, P_n$ to be a steady state it is necessary and sufficient that

$$\varphi(A_1, \dots, A_n; P_1, \dots, P_n) = 0$$

This is necessary since in a steady state the contacts of a relay have the same hindrance as is in series with the relay winding:

$$x_1 = R_1$$

or

$$x_1 \oplus R_1 = 0$$

So that

$$\varphi = \sum_1 x_1 \oplus R_1 = 0 \text{ when } x_1 = P_1, a_1 = A_1,$$

It is sufficient since

$$\varphi = 0 > P_i \oplus R_i = 0 > P_i = R_i \quad i = 1, 2, \dots, n$$

so that if the relays are held in these positions P_i long enough for fluxes to build up they will remain there.

Theorem II - For A_1, \dots, A_n to be completely oscillatory it is necessary and sufficient that

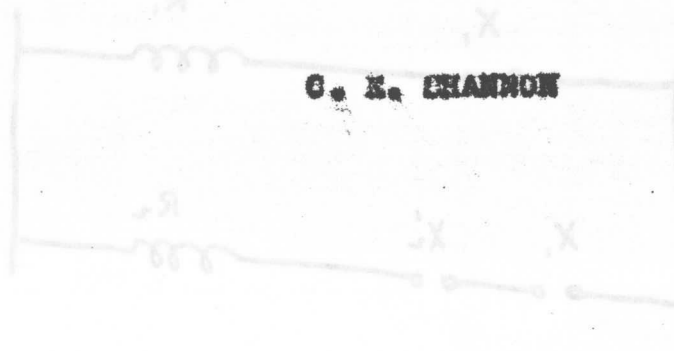
$$\varphi(A_1, \dots, A_n; x_1, \dots, x_n) = 1$$

identically in the x_i . This is necessary since otherwise there is a set of x_i , say P_i such that $\varphi = 0$ and this is a steady state by Theorem I. It is sufficient since if true then with any starting position say P_1, \dots, P_n at least one term of the sum (1) say $P_i \oplus R_i$ is equal to one, so that

$$P_i \neq R_i$$

and one or the other has to change. After some relay has changed we still have the same situation since $\varphi = 1$ so that at least one relay makes an infinite number of changes of position.

In case $\varphi(A_1, A_2, A_n, x_1, \dots, x_n)$ is a function of the x_i (not identically one or zero) the system has some steady states namely the roots of $\varphi = 0$, but for arbitrary starting conditions we cannot say what the action will be. Whether a circuit seeks out a steady state or not depends not only on the network topology as in Fig. 2, but also on relay characteristics as in Fig. 3. Here if R_1 is slow operating and R_2 very fast the circuit may chatter with both relays initially unoperated for R_2 may never stay in long enough to operate R_1 . If R_1 is fast and R_2 slow release, the system relapses into $R_1 = 0, R_2 = 1$. Hence no purely algebraic conditions can be set up to determine whether a circuit will relapse into a steady state when φ is a function of x_1, \dots, x_n .



G. E. CHANNON

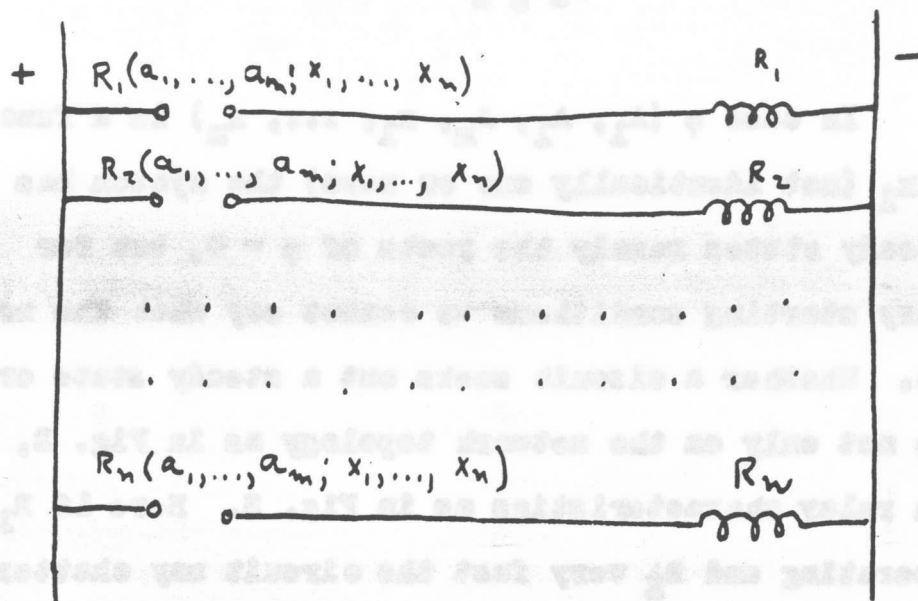


FIG. 1

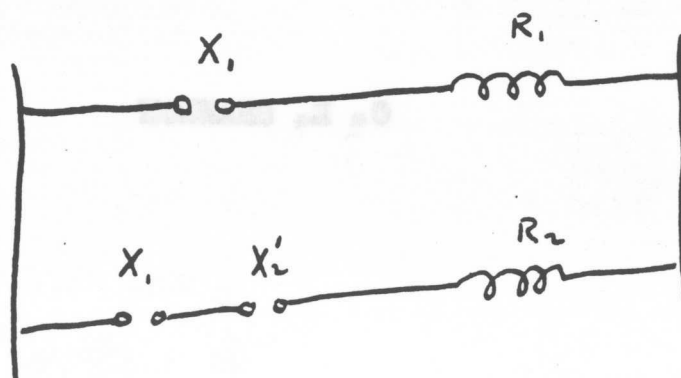


FIG. 2

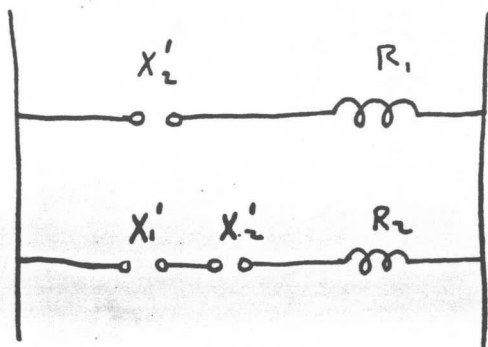


FIG. 3.

July 15, 1943

[16]

ON THE INTEGRATION OF THE
BALLISTIC EQUATIONS ON THE ABERDEEN ANALYZER

by

Professor W. Feller of Brown University and
Dr. C. E. Shannon of the Bell Telephone Laboratories

AMP REPORT NO. 28.1

APPLIED MATHEMATICS PANEL

NATIONAL DEFENSE RESEARCH COMMITTEE

RESTRICTED

This is a report on investigations made at the request of Dr. Warren Weaver (letter of December 28, 1942). Our study has been based partly on oral information received in Aberdeen (January 18, 1942) and partly on the material contained in the Report No. 319 of the Ballistic Research Laboratory ("Report on the Differential Analyzer at Aberdeen Proving Ground" by Major A. A. Bennett, December 1942). The technical set-up as described in that report will in the sequel be referred to as "present set-up". It should be clearly understood that we were not to study possible technical improvements of the analyzer as such nor to reexamine the theory underlying the differential equations. Accordingly, the present report is concerned only with an examination of the procedure of mechanical integration of the differential equations of ballistics as used at present. Furthermore, we have not considered any methods of integration other than on the differential analyzer.

Before proceeding to describe devices which might contribute to the efficiency of the analyzer we wish to summarize some negative findings, as these may render superfluous similar investigations by other persons.

a) We have carefully investigated a great number of alternative set-ups, on the differential analyzer, of the differential equations either in their present form or using various new variables. However, we have been unable to find any form superior to the method as used at present in Aberdeen

RESTRICTED

which, in our opinion, is the most efficient one.

b) We have studied the advisability of using some method of successive approximations. Such methods naturally present themselves since one should expect them to reduce the ranges of the variables involved and thus increase the accuracy. However, a closer study will show that it is almost invariably necessary to subtract, on the analyzer, two large quantities which are themselves independently obtained on the analyzer. This, of course, nullifies the desired effect of reducing the ranges. Various possibilities have been studied and, among them, the possibility of starting with the vacuum trajectories and integrating the difference between them and the actual trajectories. Again we were unable to find a method which would appear superior to the present set-up. It will be noted, however, that the modification of the latter suggested below, can in some sense be interpreted as the first step in method of successive approximations.

c) Several perturbation methods and expansions according to various parameters have been tried paying special attention to methods suggested in the newest Russian literature. None of these methods seem appropriate for the analyzer.

Coming to the less negative part of this report we remark that an adequate theory of errors of the differential analyzer is not available at present. However, simple theoretical considerations based on experience gathered at M.I.T. make it appear that a very considerable part of the total error is due

RESTRICTED

to the slippage of parts of integrators; the other main sources of error are backlash and, perhaps even more, inaccuracies in the following mechanism for the input and vector tables. It seems therefore possible to achieve a gain in accuracy by reducing the range of the variables in the integrators, even though this may necessitate the introduction of new adders and gears. The following recommendations are based on this assumption. We proceed step by step starting with the simplest case.

Recommendations.

1) Consider, to begin with, the horizontal displacement x . Obviously dx/dt will range from its maximum \dot{x}_0 at the beginning to some fraction of it, say $q\dot{x}_0$, at the end. Accordingly, when integrating in the usual form

$$(1) \quad x = \int \dot{x} dt,$$

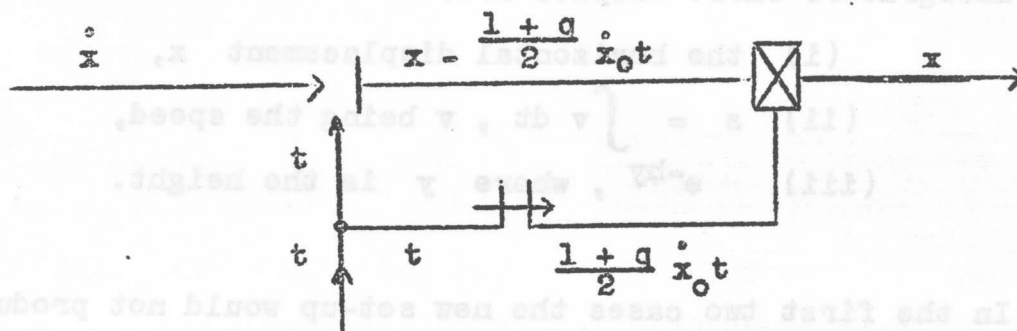
the integrand ranges from $q\dot{x}_0$ to \dot{x}_0 . Now this means that only a fraction $\frac{1+q}{2}$ of the total range of the integrator disc is used even if we suppose that the scale factor has been chosen in the best way (so that the rim of the integrator disc is used for values of \dot{x} near \dot{x}_0). If, instead, we write

$$(2) \quad x - \frac{1+q}{2} \dot{x}_0 t = \int (\dot{x} - \frac{1+q}{2} \dot{x}_0) dt,$$

RESTRICTED

the integrand will range from its maximum $\frac{1-q}{2} \dot{x}_0$ to its minimum $-\frac{1-q}{2} \dot{x}_0$. This allows one to use a scale factor $\frac{2}{1-q}$ times as large as in the set-up (1) and to utilize the entire integrator disc. This, of course, means a considerable gain.

Now the constant $\frac{1+q}{2} \dot{x}_0$ in the integral in (2) appears only as an initial displacement. It is therefore seen that the realization of the proposed set-up (2) requires, as compared with the customary set-up (1), an additional gear (to produce $\frac{1+q}{2} \dot{x}_0 t$) and an adder. The following figure shows the simplest mechanization.



It goes without saying that the gear ratio does not need to be exactly $\frac{1+q}{2} \dot{x}_0$: any number near the middle of the range of the integrand will do the same services.

If used to its fullest extent, the system as described changes a previously positive variable into one taking on also negative values. Although only one change of sign is introduced, this will introduce some new backlash. Now, if instead of (2) we mechanize

$$(3) \quad x - q\dot{x}_0 t = \int (x - q\dot{x}_0) dt,$$

RESTRICTED

the new integrand does not change sign, and no new backlash is introduced. On the other hand, the optimum scale factor for (3) is only $\frac{1}{1-q}$ times that for (1), that is to say half the scale factor for (2). We conclude that with proper corrections for backlash the set-up (2) should prove best. However, if enough frontlash units are not available at Aberdeen, the set-up (3) may be tried with advantage.

2) A similar device can obviously be used wherever the range of the integrand does not utilize the integrator disc to its fullest extent. This is true for almost all integrators whose outputs are:

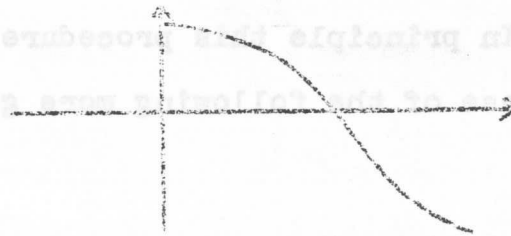
- (i) the horizontal displacement x ,
- (ii) $s = \int v \, dt$, v being the speed,
- (iii) e^{-hy} , where y is the height.

In the first two cases the new set-up would not produce any additional loading since the integrators are driven by the independent variable-motor. In other cases an additional loading would ensue which may have to be compensated by the use of a larger scale factor on the t-shaft; this would indirectly slow down the machine. Whether this will have to be done is impossible to predict theoretically. Should it prove necessary, it would be for the user to decide whether the gain in accuracy is worth the loss in speed.

3) If the above described device should prove in-

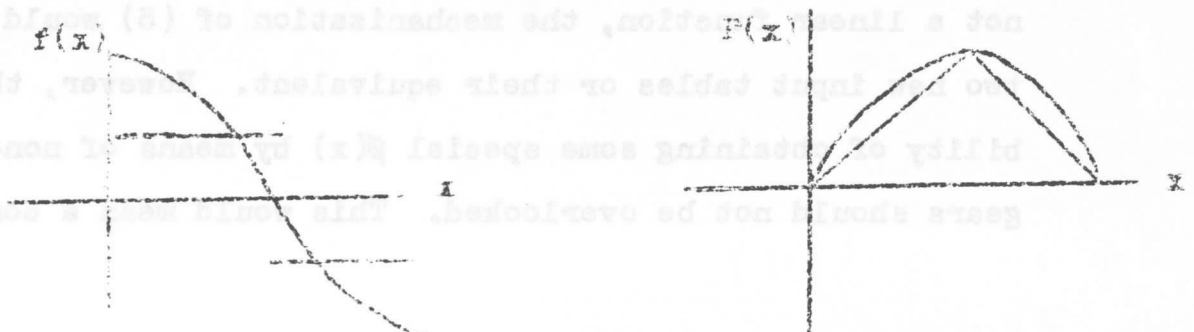
RESTRICTED

insufficient to reduce the error to the desired tolerance, the following improvement may work, although at the expense of considerable manual work and loss of speed. The process of integration may be stopped at convenient moments and the procedure as described above may be used for each of the sub-intervals. Consider, for example, an integrand of the form indicated in the figure (such as $\cos x$)



Here even the usual procedure of integration utilizes the entire range of the integrator disc and no gain can be achieved by means of the device as described above. However, the integrand may conveniently be treated by a double application of this device splitting the interval of integration into two parts.

In other words, instead of a given function $f(x)$ we integrate the difference between $f(x)$ and a step-function. The output of the integrator is no longer $F(x) = \int f(x) dx$, but the difference between $F(x)$ and a triangular (or "roof") function.



Similarly, with a convenient subdivision we may use any step-function for the integrand and the corresponding polygonal line for the integral.

This procedure obviously requires resetting the integrator in question and changing one gear ratio each time the machine is stopped. On the other hand, the increase of the scale factor is roughly proportional to the number of subintervals.

4) In principle this procedure may be looked upon as a special case of the following more general method. Instead of

$$(4) \quad w(x) = \int y \, dx$$

write

$$(5) \quad w(x) + \phi(x) = \int (y + \phi') \, dx,$$

where $\phi(x)$ is an arbitrary function and $\phi'(x)$ its derivative.

In practice, of course, $\phi(x)$ should be chosen so as to render the maximum of $|y + \phi'|$ as small as possible in order to increase the scale factor on the integrator. Now if $\phi(x)$ is not a linear function, the mechanization of (5) would require two new input tables or their equivalent. However, the possibility of obtaining some special $\phi(x)$ by means of non-circular gears should not be overlooked. This would mean a considerable

improvement of the linear method.

5) We have been asked by Dr. Dederick to consider whether it would be advantageous to generate e^{-hy} from an input table (instead of by integration, as at present). The foregoing remarks contain an answer to this question. It is not difficult to see that the present method of obtaining the function by integration is more efficient. It would probably become even more so if the recommendation 2) were put into effect.

6) Although it is in no direct connection with the subject of this report, we enclose an Appendix describing a simplified method for computing gear ratios. This method is based on previous experience (of one of us) at M.I.T. and may prove useful in connection with ballistic work on the Aberdeen Analyzer.

Brown University, Providence, R.I.
and
Bell Telephone Laboratories, N.Y.

May 27, 1943.

W. Feller

C.E. Shannon

RESTRICTED

A METHOD OF DETERMINING GEAR RATIOS

In this appendix a simplified method of determining gear ratios for an analyzer set up will be described which was used for some time on the M.I.T. analyzer and proved in general to be considerably faster and easier to change than the original method of equalities and inequalities. The method may be briefly outlined as follows:

1. Draw the set up with an unknown gear ratio in each shaft of limited displacement. An unspecified ratio is also placed in the two inputs of each adder.
2. Calculate an approximate scale factor on the independent variable to give the expected time of solution at the average rate at which it turns. Choose an exact scale factor near this approximate one which is a "round figure" in terms of obtainable gear ratios - i.e., factorable into a small number of simple rationals.
3. Choose in the same way scale factors for all shafts of limited displacement - integrator inputs and function table inputs, and outputs - so as not to exceed their limits with expected displacements.
4. This fixes, by division, and from the integrating factor of the integrators, the scale factors and gear ratios of all shafts except those containing adders. In the case of adders the input shaft with smallest scale factor fixes the scale factor of the adder, the other input being geared down to the same scale factor. The output gear in the adder is then fixed.
5. The set up is then inspected to see that no integrators or other parts are too heavily loaded. If they are, reduction gears are transferred from inputs to outputs to reduce loads when possible, otherwise the scale factor on the independent variable is increased.

In case the ratios come out too complicated different scale factors are chosen in Step 3. With a little practice and foresight, however, it is possible to obtain suitable ratios on the first trial.

RESTRICTED

Two New Circuits for Alternate Pulse Counting

The well known W-Z relay circuit is shown in Fig. 1. A is a pulsing contact which is alternately opened and closed. Indicating closure of contacts by 0 and openness by 1 and for relays 0 for operated (up) and 1 for unoperated (down) the circuit goes through the following periodic cycle of operation:

A	W	Z
1	1	1
0	0	1
1	0	0
0	1	0
1	1	1

Thus one complete cycle requires two complete pulses on A.

This note describes two apparently new circuits which perform the same function. These are shown in Fig. 2 and Fig. 3. The operating cycles for these are:

Fig. 2

A	W	Z
1	0	1
0	0	0
1	1	0
0	1	1
1	0	1

Fig. 3

A	W	Z
1	1	1
0	0	1
1	0	0
0	1	0
1	1	1

These three circuits may be compared with regard to the number of elements required as follows:

	Relays	Contacts	Resistances
Figure 1	2	1 continuity, 1 transfer	2
Figure 2	2	2 continuity, 1 break	1
Figure 3	2	2 transfer, 1 make	1

In Fig. 3 the resistance is theoretically superfluous; if the transfer elements could be trusted never to be shorted it could be omitted, but in practice would be necessary to avoid shorts when the relays were being adjusted. Figs. 2 and 3 are essentially duals, and 3 was obtained from 2 by the duality theorem.

In Fig. 2 it may be noted that the two relays are ~~up~~ ^{down} when A is closed, while in the standard circuit they are both ~~down~~ ^{up} when A is open. This might be desirable in some applications. Fig. 3 has the possible disadvantage that both ends of the pulsing contact A are connected into the circuit, while in 1 and 2 one end can be grounded.

C. E. SHANNON

Att.
Figs. 1, 2, 3

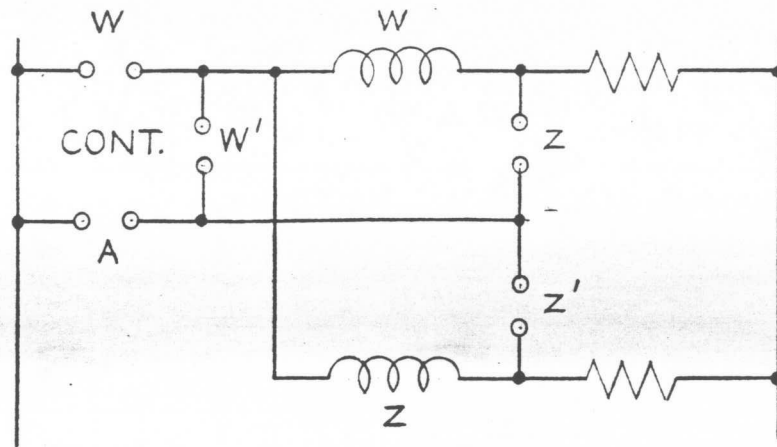


FIG. 1

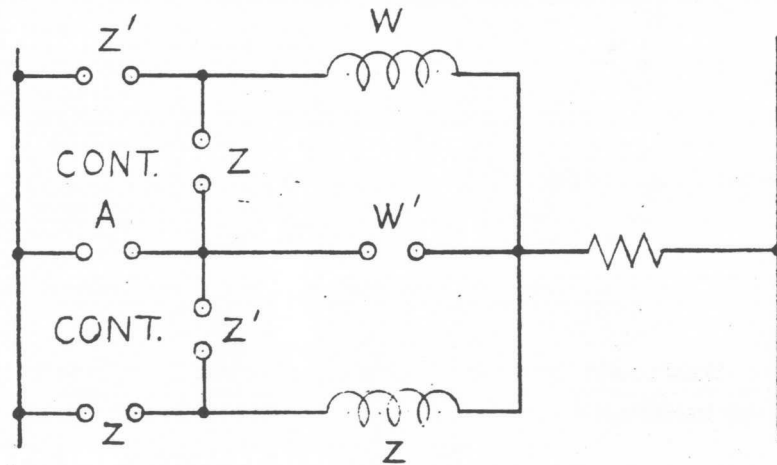


FIG. 2

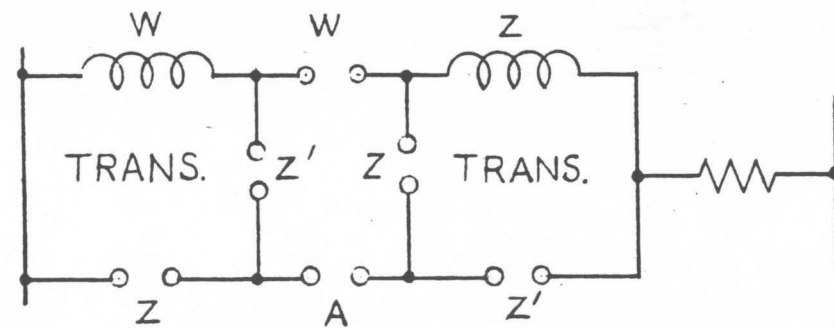


FIG. 3

DR.		CH.	TITLE
L.			
DR.		ENG. C. E. SHANNON	SCALE
L.			
BELL TELEPHONE LABORATORIES, INC., NEW YORK			

Counting Up or Down with Pulse Counters

With binary counters of either relay or electronic type it is possible by simple modification to count both up and down. Suppose the largest number that can be registered is L . Defining the complement of any number $N < L$ by $L-N = N'$ we note that subtracting a number N from N is equivalent to adding N to its complement $N-B = L-N'-B = L-(N'+B)$. Thus if in a binary counter we take the complement of a reading (which means locking up the relay which are down and vice-versa in the ~~relay~~ ^{relay} case, and putting out the tubes which are conducting and vice-versa in the electronic case) and then let the counter continue add the number of pulses in question, and finally take the complement again, we have subtracted the number. Actually however, this process can be done simply by transferring the carryover leads to the opposite digit (tube or relay). In the relay case this amounts to a transfer contact between each adjacent pair of digits, and an additional make contact. In the electronic case the carryover leads go from the "zero" tube plate to grids on the next stage. Here we could insert an electronic transfer contact, as shown, for example in Figure 1. When we wish to add, the common control leads for "add" is given cutoff voltage, the "subtract" lead a large negative voltage. A positive impulse on the "one" plate of a stage then causes one side of the double triode to conduct giving a negative impulse to the next grids for a carryover. For subtraction the voltages on the control leads are reversed and carryover occurs when the "zero" plate voltage increases i.e., when this tube goes out.

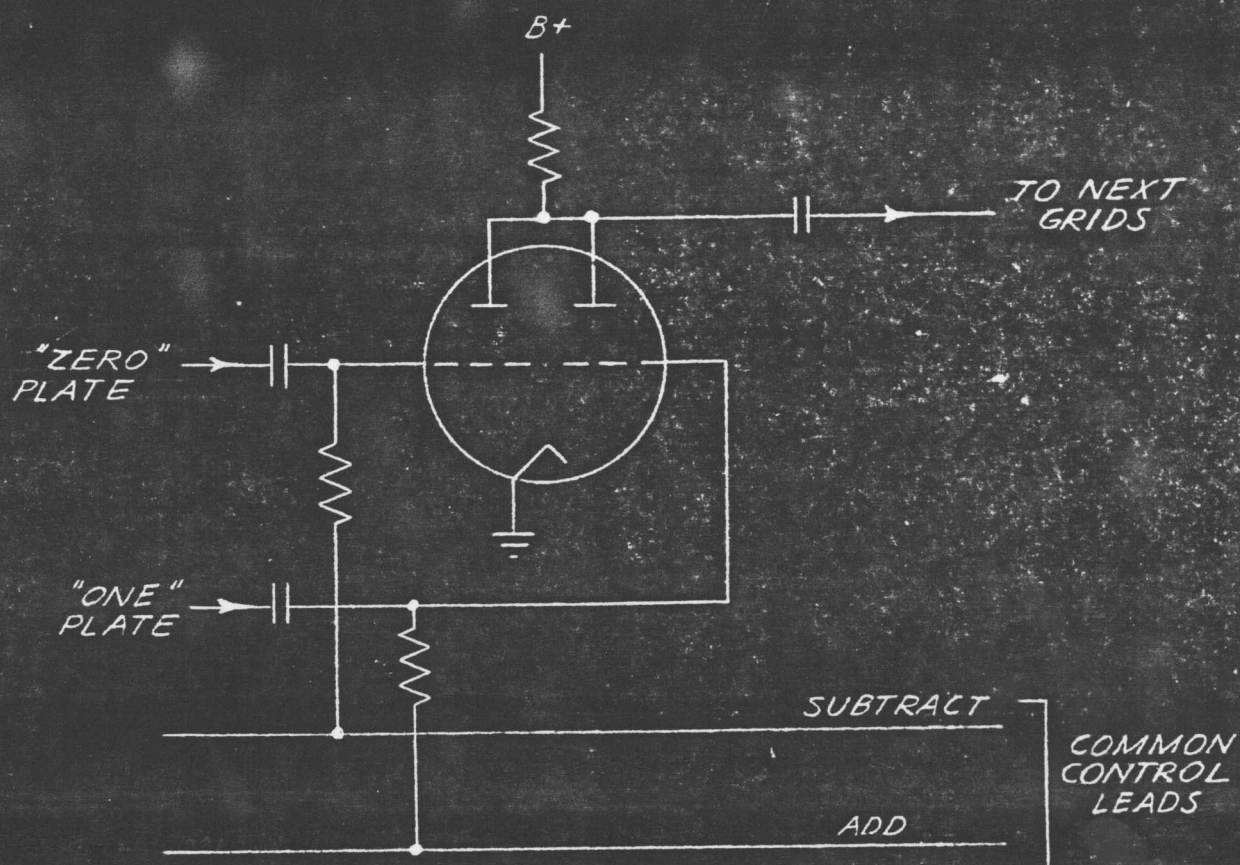


FIG -1

ISSUE 1 5-31-44

APPL.	DR. R/V	CH.
	ENG. CE SHANNON	

TITLE
SCALE
BELL TELEPHONE LABORATORIES, INC., NEW YORK

(U11)

E-1751 (9-43)

[21]

COVER SHEET FOR TECHNICAL MEMORANDA
RESEARCH DEPARTMENT

SUBJECT: Circuits for a P.C.M. Transmitter and Receiver -
Case 20878

ROUTING:

- 1 - S.A.S., H.W.B., H.F.
2 - CASE FILES
3 - ~~R. Bown~~
4 - G.W. Gilman
5 - H.W. Bode
6 - A.G. Jensen
7 - W.M. Goodall
8 - E. Peterson
9 - H.S. Black
10 - W.F. Simpson - Patent Dept.
11 - J.R. Pierce
12 - R.L. Dietzold
13 - C.B. Feldman
14 - W.T. Wintringham
15 - F.B. Llewellyn
16 - C.H. Elmendorf
17 - B.M. Oliver
18 - C.E. Shannon

MM-- 44-110-37
DATE June 1, 1944
AUTHOR s C.E. Shannon and
INDEX NO. B.M. Oliver

~~ABSTRACT~~

ABSTRACT

Circuits are described for a P.C.M. transmitter and receiver. The transmitter operates on the principle of counting in the binary system the number of quanta of charge required to nullify the sampled voltage.

GROUP 4
Downgraded at 3-year intervals;
declassified after 12 years

MM-44-110-37

June 1, 1944

MEMORANDUM FOR FILE

The circuits shown in the present memorandum are intended to fill the boxes of the block functional designs for a PCM transmitter and receiver shown in Fig. 6 of a December 1943 memorandum (MM-43-110-43). The transmitter functional diagram is shown here as Fig. 1 and the general operation is as follows. The incoming signal is sampled periodically by closing the electronic switch 1 with periodic impulses from the timer. This charges condenser C to the sampled voltage and the electronic switch opens after each impulse isolating the condenser from the signal. The existence of a voltage across the condenser causes the comparator to close electronic switch 2 which allows pulses of charge to feed into the condenser from the pulse generator, discharging the condenser. The number of these pulses is counted in the binary system by the binary counter and when the condenser is reduced to a reference voltage, the comparator opens electronic switch 2. Near the end of the sampling period the binary counter is connected to the distributor which registers the binary number counted, and the counter is then reset to zero; both of these operations controlled by impulses from the timer. The distributor then sends a series of pulses or not down the output line according as the binary digits are 1 or 0. These digits are sent in reverse order, the least important being sent first, to tie in with the contemplated receiver circuit.

The specific circuits are shown in Figs. 2 to 8, and detailed descriptions of their operation follow.

Fig. 2 shows the electronic switch 1 which charges the condenser C to the signal voltage at the sampling times. The signal wave is biased up so that its minimum value is slightly positive, and impressed on terminal 1 as a voltage; i.e. the signal source as seen from terminal 1 is assumed to be of low impedance. The timer, at the sampling time puts a positive pulse on terminal 2, which is inverted by the triode to give a negative pulse on the pentode control grid. This causes the pentode which was previously conducting to cut off. Before the pulse condenser C had a small minimum positive charge and neither diode was conducting since the plates were held at a low positive potential by the pentode current. As the

pentode cuts off, the diode plates swing positive and the right hand diode starts to conduct charging the condenser. As this condenser voltage builds up exponentially the voltage on the diode plates also increases positively until it reaches the signal voltage and at that instant the left hand diode starts to conduct. The voltage stops rising at this point since the plates are now essentially short circuited to the low impedance signal source. This all occurs during the timing pulse, and at the end of this pulse the pentode again starts conducting dropping the diode plates to a small positive voltage, less than the minimum signal voltage, and isolating the condenser.

Fig. 3 shows a standard multi-vibrator circuit for giving a series of square pulses. The coil condenser cross connection of plates to grids causes the grid transient to be a cosine curve which crosses the cut off grid voltage at a time determined essentially by the LC product and independent of amplitude changes due to variations in plate supply, etc. As this point determines the period of oscillation, the oscillator has good frequency stability. The output appears on terminal 6 as a square wave.

Fig. 4 is the comparator, which is actually only a differential amplifier with sufficient gain so that the granularity voltage applied to the input is capable of driving the amplifier from saturation in one direction to saturation in the other. The input is the voltage on condenser C which immediately after a sampling instant, will be at the sampled signal voltage. This voltage starts decreasing by steps as the condenser is discharged and when the condenser voltage applied to terminal 3 moves down the step which crosses the differential amplifier threshold, the amplifier swings from saturation with output terminal 5 at nearly zero voltage to a high negative voltage.

The electronic switch 2 is shown in Fig. 5. This circuit sends units of charge into the condenser through terminal 3 under the control of the comparator output coming in on terminal 5. The multi-vibrator output is connected to terminal 6 and the output of the multi-grid tube will be a square wave when 5 is positive, which ceases when the comparator swings to the other saturation point driving the voltage on 5 in the negative direction. The double diode connection gives a pump action. When the plate voltage of the multi-grid tube increases to the upper part of the square wave, the charge flows into the condenser from terminal 4 through the left diode. During the lower part of this wave

the condenser discharges through the right diode out into the condenser C, via terminal 3. As this causes the potential of 3 to decrease gradually down a step function, it is necessary for the input voltage at 4 to decrease similarly; otherwise the difference in voltage between 3 and 4 would cause the size of quanta to decrease gradually. This lowering of the voltage on 4 is accomplished by a cathode follower arrangement on the first cathodes in the comparator, which follow the step voltage down.

The binary counter is shown in Fig. 6. The descending step voltage which appears on condenser C is applied to the input of this circuit through terminal 3. The input resistance condenser combination serves as a differentiating circuit (the time constant fairly small compared to the time between steps) so that the voltage applied to the first grid of the double triode consists of a series of negative spikes. The double triode is simply a two stage resistance coupled amplifier, and its output feeds the binary counter digit tubes. This circuit is of standard type with two pentodes in each stage and there are two stable points for each stage, one with the upper tube cut off and the lower tube conducting, and the other, the converse situation. A negative impulse from a preceding stage applied through the coupling condensers changes the state from the previous stable condition to the opposite one. This impulse is applied symmetrically to both suppressors, but the condenser across the cathode resistances, charged in one direction from the previous state, biases the choice of the next state toward the opposite one. The control grids of the "zero" tubes (the upper row which are conducting when the corresponding binary digits are zero) are connected to a common control lead which is used to reset the reading to zero after the reading is registered by the distributor. This is accomplished by a negative impulse from the timer. The outputs to the distributor are taken off the plates of the "unit" tubes.

The distributor is shown in Fig. 7. After the number of quanta of charge has been counted in the binary counter, the leads 11, 12, 13, 14, 15 will have either low positive voltages or B+, according as the corresponding digit is one or zero. The grids of the left triode, will then be either negative or positive from the potentiometer action to the negative voltage C-. To register the counter reading, a positive pulse from the timer is applied to the control grid of the common pentode allowing it to conduct and pulling the cathode of the left triode and the diode in all stages negatively. If a digit is zero, the potential of the cathodes in that stage stops at a positive value due to current through the triode and the diode does not conduct. If the digit is one the cathodes are pulled negative and the corresponding

condenser C_0 is discharged through the diode and pentode. At the end of the registering pulse, the cathodes go positive again, isolating each C_0 , with the digit registered as presence or absence of charge. The reading is taken off the series of condensers C_0 in sequence by positive pulses from the timer on leads 21, 22, 23, 24, 25. These pulses allow the right hand triodes to conduct and each C_0 in turn to charge through the output lead, leaving them in the normal state (at a voltage about equal to the pulse voltage). If the digit is "zero" no charge of C_0 from the output lead occurs. Thus negative pulses appear on the output when and only when the registered digits are one.

The timer system is shown in Fig. 8. An oscillator which may be synchronized subharmonically with the pulse generating multi-vibrator, operates at the sampling frequency. This passes through the clipper amplifier to give a square wave, which is differentiated to give alternating positive and negative spikes. A second clipper amplifier eliminates the negative spikes and makes the positive ones rectangular. These short rectangular pulses are fed into a delay line terminated in its characteristic impedance. The timing pulses needed for the various circuit functions are tapped off at the appropriate places as indicated. A synchronizing pulse may also be taken off the same delay line.

Fig. 9 shows the receiver circuit. The signal passes through the clipping amplifier which is adjusted to give a saturation voltage on the output if a pulse is present and none if absent. This output is applied to the grid of a multigrid pentode, whose other control grid is given positive gating pulses at the center of the digit intervals. These gating pulses allow the pentode to conduct if a pulse is present and the plate current is then independent of the plate voltage (providing this stays within certain limits) so that if a pulse is present, a fixed amount of charge (equal to the length of the gate times the pentode current) flows onto the condenser. The time constant of the RC system (including the pentode load resistance) is adjusted to allow the voltage to restore itself halfway toward the equilibrium value in the time from one digit to the next, so that after all pulses have been collected on the condenser, the charge contributions of the first, second, third etc. have decayed by factors of $\frac{1}{2^4}$, $\frac{1}{2^3}$, $\frac{1}{2^2}$, $\frac{1}{2}$, 1. At this time a positive gating pulse is put on the grid of the second pentode, allowing the condenser to discharge rapidly into the low pass filter. The timer system can be realized with the systems shown in either Fig. 10 or Fig. 11.

C. E. SHANNON
B. M. OLIVER

Att.
Figs. 1 to 11

CONFIDENTIAL

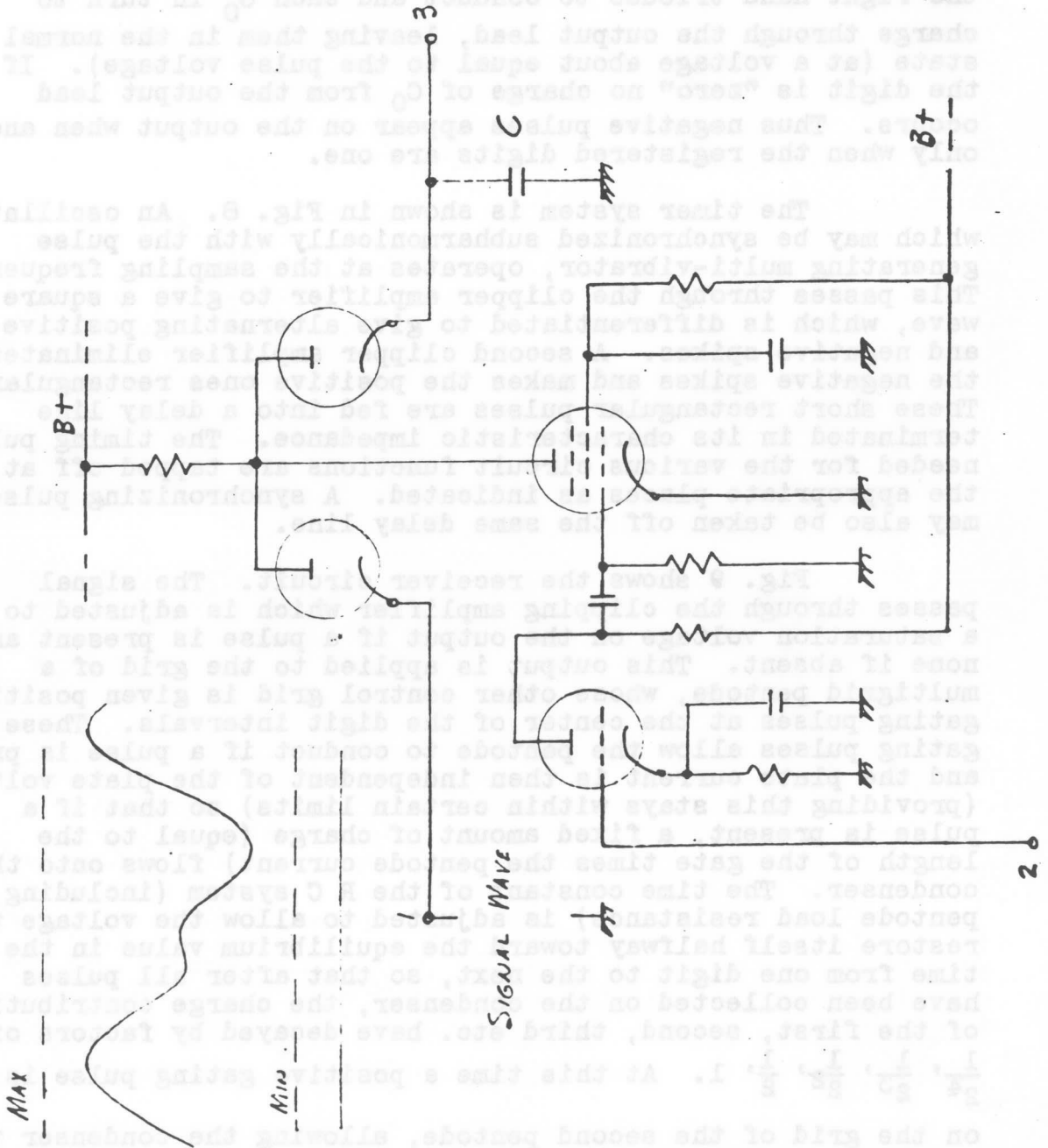


Fig 2

C. E. SHANNON
B. M. OLIVER

Fig. 1 to 11

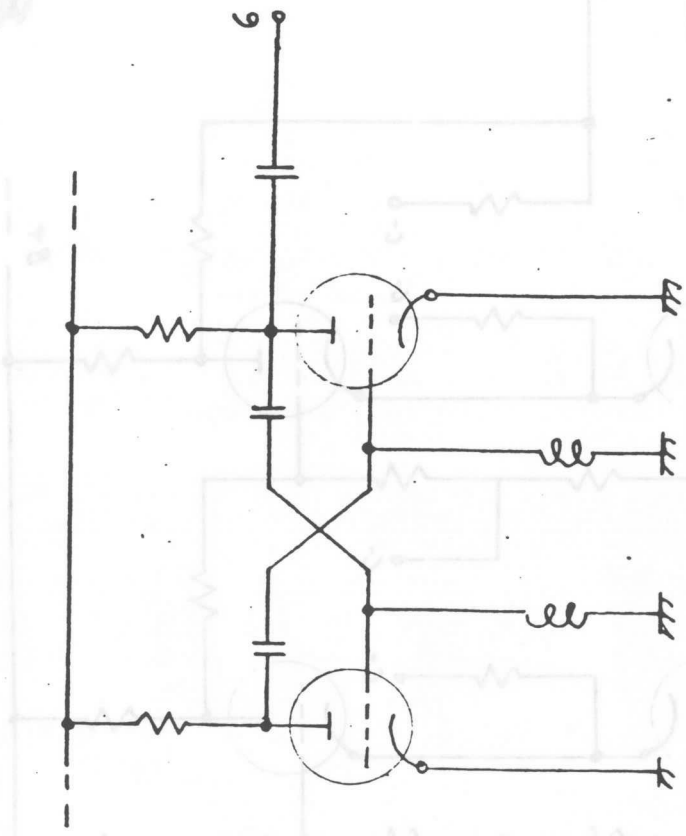


FIG. 3

Fig. 4

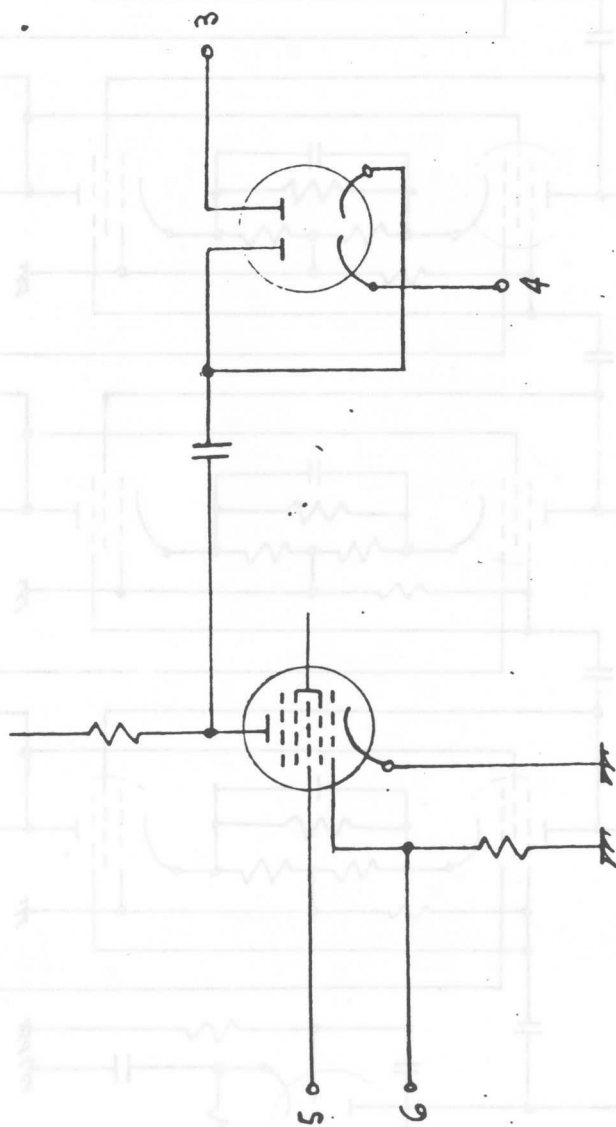
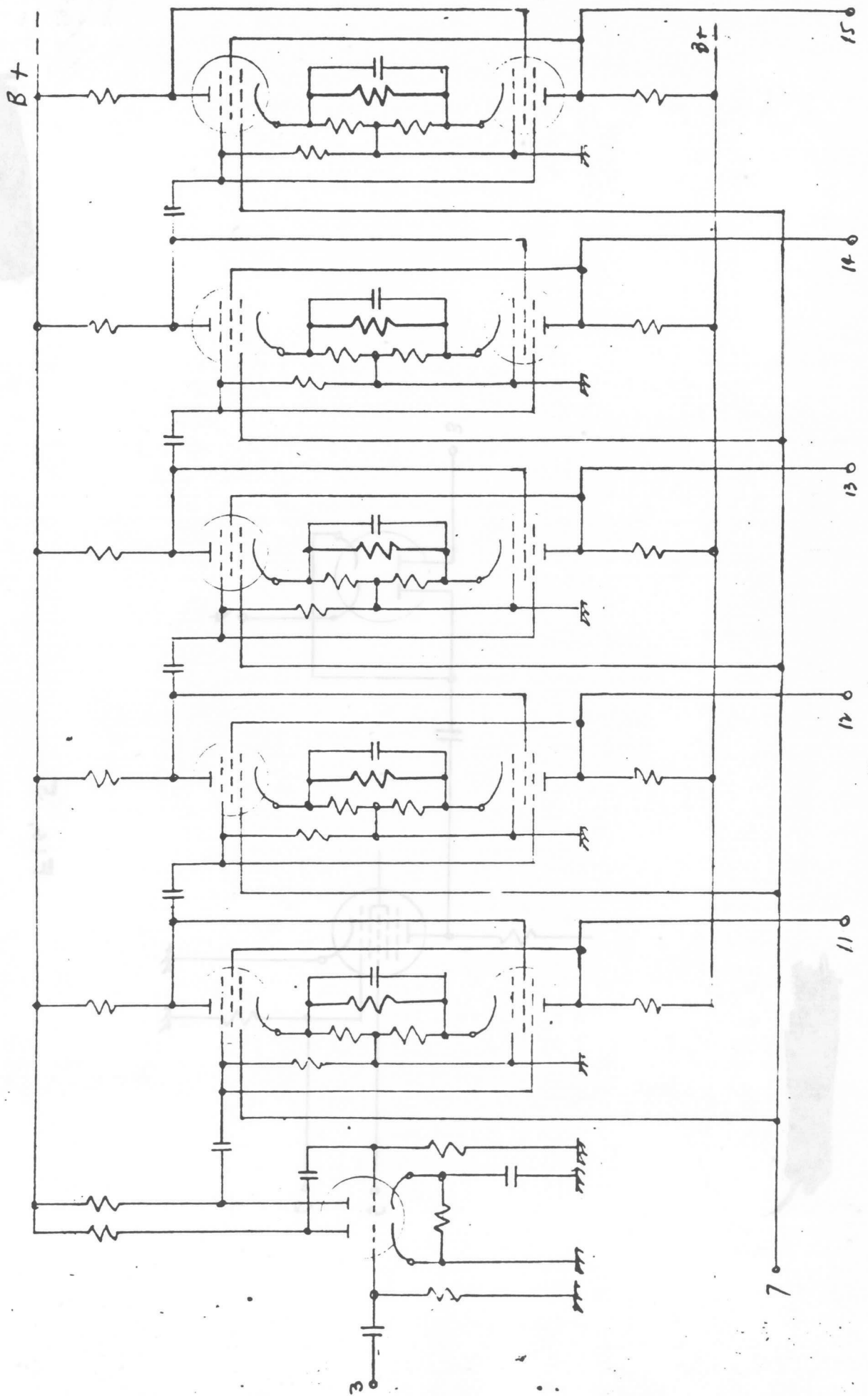
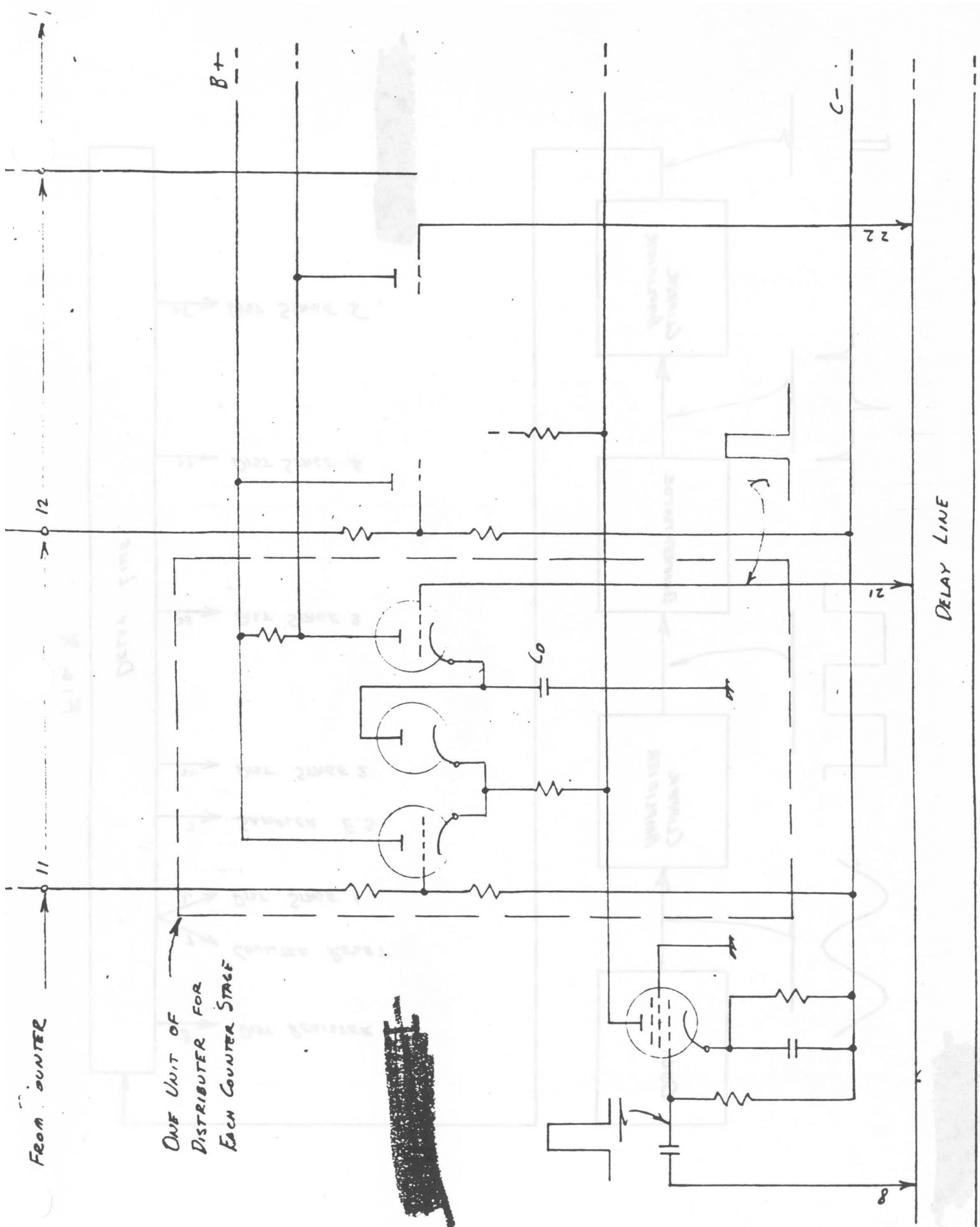


FIG. 5

~~CONFIDENTIAL~~





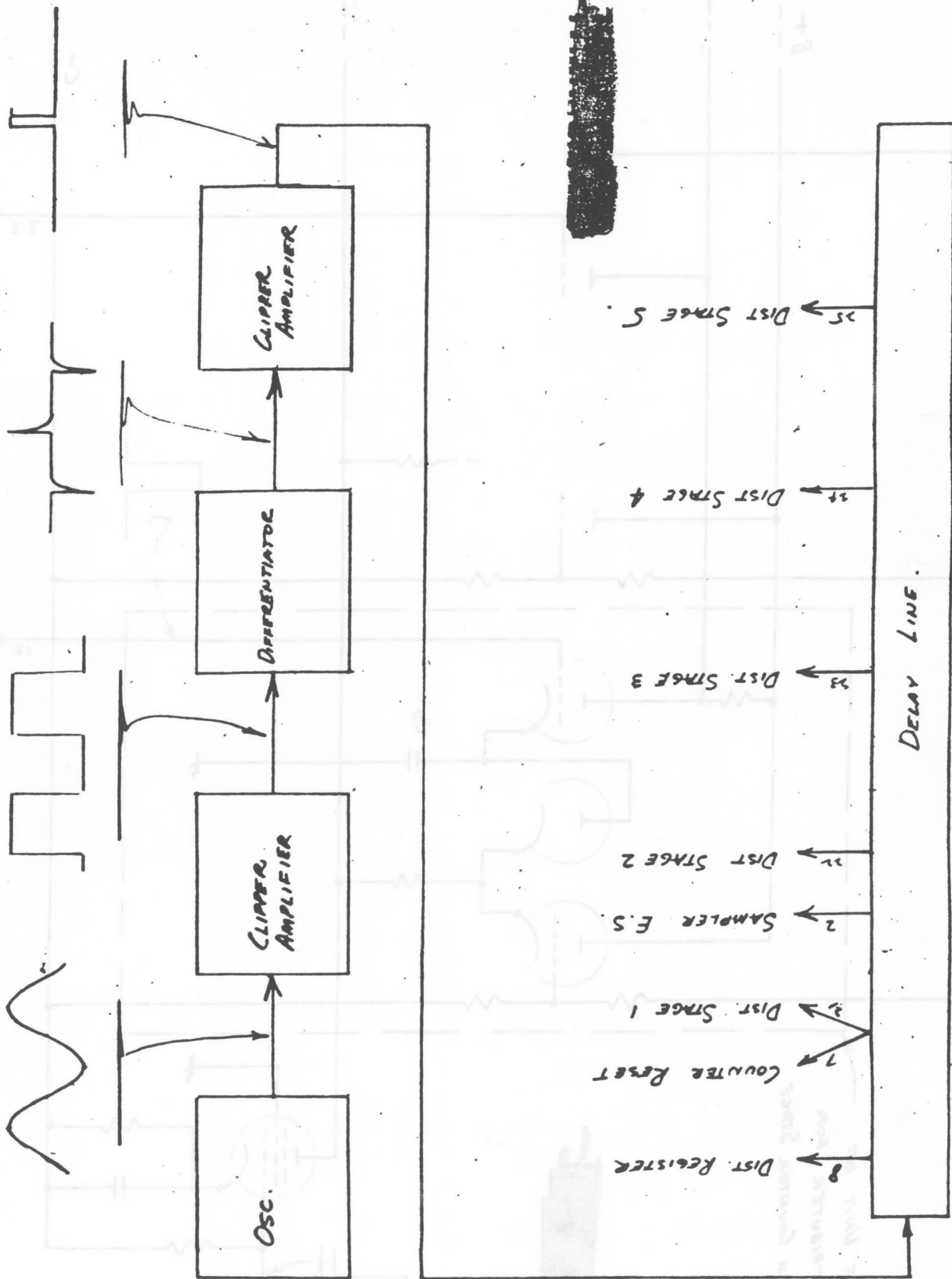


FIG. 2

FIG -9

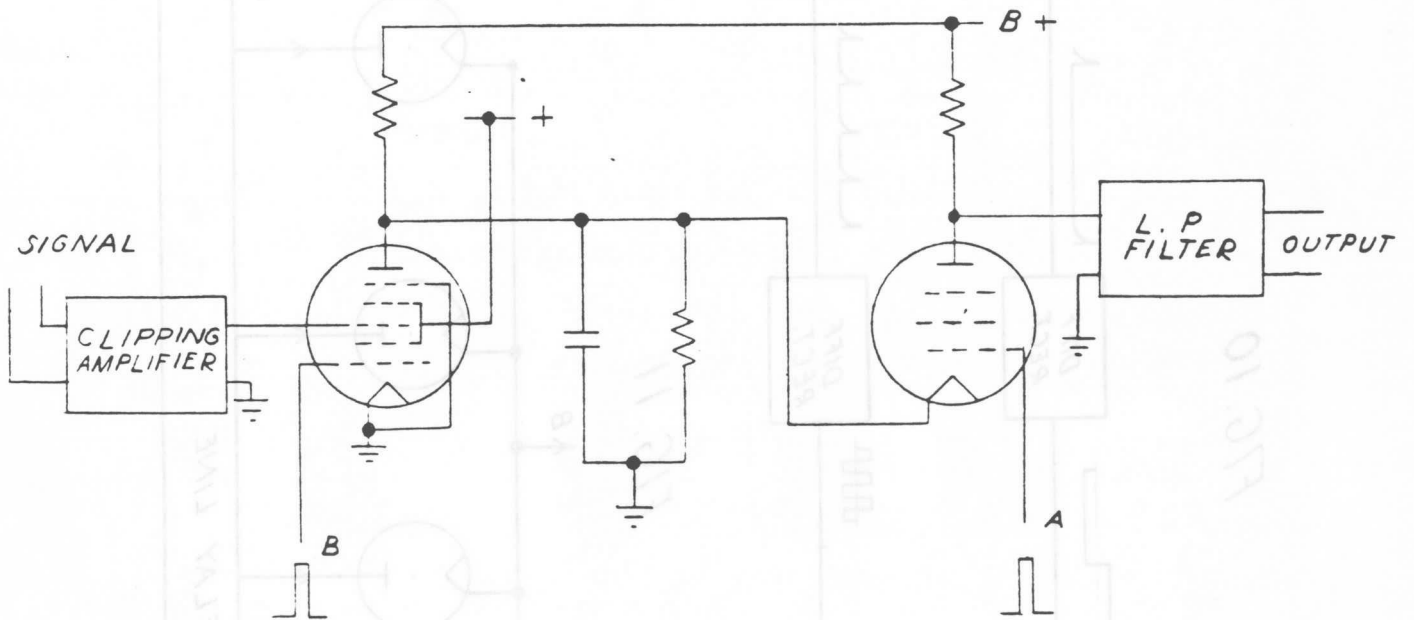


FIG. 10

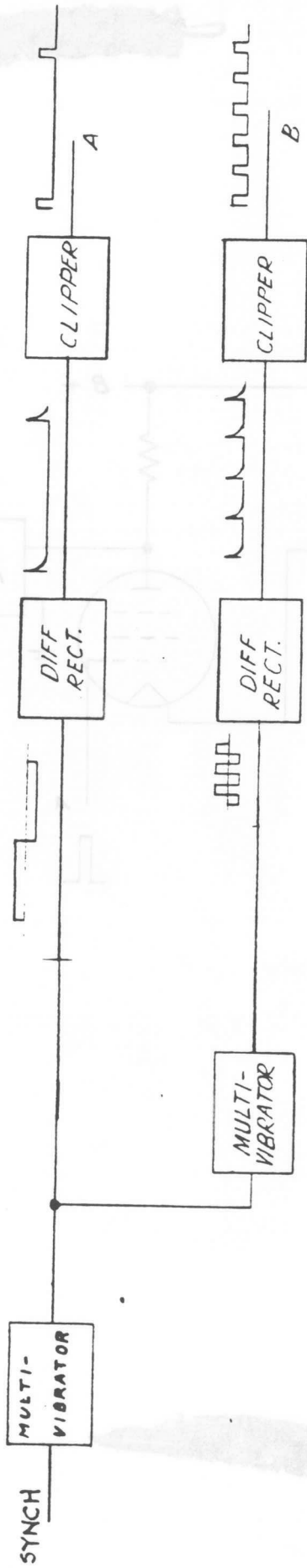
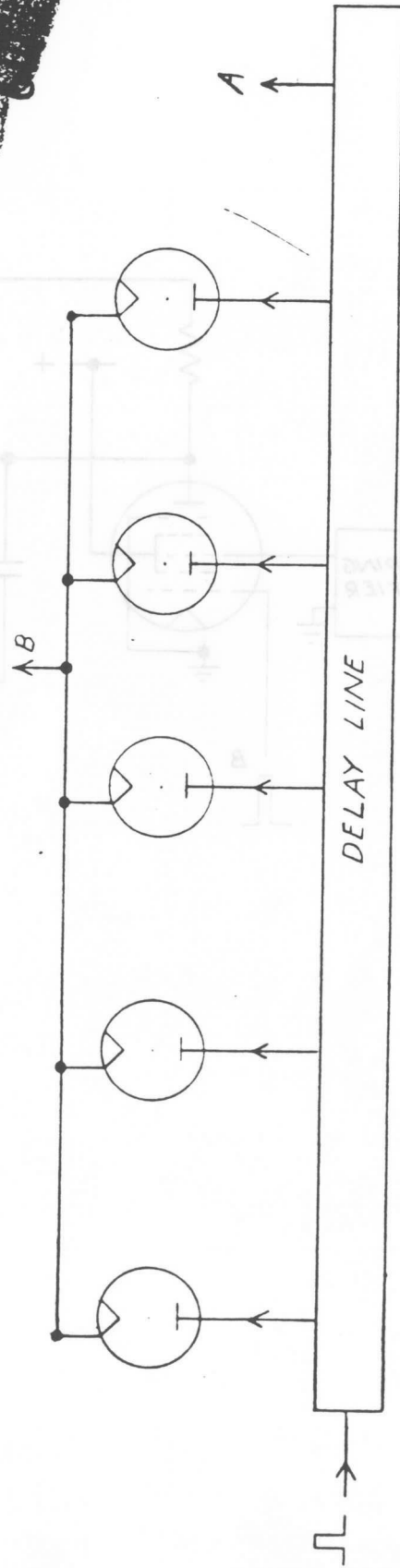


FIG. 11



E5 | 111

August 4, 1944

Pulse Shape to Minimize Band Width with Nonoverlapping Pulses

We consider the problem of shaping pulses $\varphi(t)$ which are zero outside $-L, L$ in such a way as to minimize the band width of the power spectrum of the ensemble of functions formed by sending a sequence of the functions $\varphi(t)$ and 0, with spacing of $2L$, the probabilities of either being $1/2$.

First a theorem will be proved on the spectrum of such ensembles of functions.

Theorem: Let an ensemble of functions be defined by

$$f(t) = \sum_{n=-\infty}^{\infty} a_n \varphi(t+2nL)$$

where the a_n are chosen independently and are equally likely to be one or zero. The power spectrum of $f(t)$ then consists of two parts, a point spectrum consisting of the spectrum of $\frac{1}{4} \sum \varphi(t+2nL)$, i.e. the spectrum of $\varphi(t)$ repeated, and a continuous part consisting of the energy spectrum of $\frac{1}{4} \varphi(t)$.

Consider the autocorrelation of $f(t)$

$$\begin{aligned} A(\lambda) &= \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T f(t) f(t+\lambda) dt \\ &= \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T \sum a_n \varphi(t+2nL) \sum a_m \varphi(t+2mL+\lambda) dt \end{aligned}$$

The integrand can be written

$$\begin{aligned} & \sum_{m \neq n} a_m a_n \varphi(t+2\pi n) \varphi(t+2\pi m+\lambda) \\ & + \frac{1}{2} \sum a_m \varphi(t+2\pi m) \varphi(t+2\pi m+\lambda) \\ & + \frac{1}{2} \sum a_m^2 \varphi(t-2\pi m) \varphi(t+2\pi m+\lambda) \end{aligned}$$

When we average, the sum of the first two parts gives the autocorrelation of the function $\frac{1}{2} \sum \varphi(t+2\pi n)$ since the coefficients $a_m a_n$ ($m \neq n$) have one chance in four of being both equal to one, and in the second term $\frac{a_m^2}{2}$ has the same mean value.

The last term in the limit reduces to

$$\frac{1}{4} \int_{-\infty}^{\infty} \varphi(t) \varphi(t+\lambda) dt$$

since the division by $2T$ compensates for the number of terms.

These two parts give the discrete and continuous parts of the spectrum, the first being the autocorrelation of $\varphi(t)$ repeated and the second giving the energy spectrum of $\varphi(t)$ when we take the transform.

In case $\varphi(t) = 0$ outside $-\pi, \pi$, the discrete part has power at $s = 0, 1, 2, 3, \dots$ amounting to

$$P_1 = \frac{1}{4} [a_0^2 + (a_1^2 + b_1^2) + \dots]$$

where

$$\varphi(t) = a_0 + \sum a_n \cos nt + \sum b_n \sin nt.$$

Suppose we wish to shape $\varphi(t)$ lying within $-1, 1$ in such a way as to minimize the band spread of the spectrum as measured by

$$W = \int_{-\infty}^{\infty} \omega^2 P(\omega) d\omega.$$

The contributions of the two parts of the spectra can be added, and that from the discrete part is

$$W_1 = \frac{1}{4} [(a_1^2 + b_1^2) 1^2 + (a_2^2 + b_2^2) 2^2 + \dots]$$

For the continuous part using the theorem that the $\int_{-\infty}^{\infty} \omega^2 F^2(\omega) d\omega = \int_{-1}^1 [f'(t)]^2 dt$ where $f(t)$ and $F(\omega)$ are Fourier transforms we have

$$W_2 = \frac{1}{2} \int_{-1}^1 \varphi'(t)^2 dt = \frac{1}{4} [(a_1^2 + b_1^2) 1^2 + (a_2^2 + b_2^2) 2^2 + \dots]$$

i.e., the same as the discrete contribution. The total W is therefore

$$W = \frac{1}{2} [(a_1^2 + b_1^2) 1^2 + (a_2^2 + b_2^2) 2^2 + \dots]$$

To minimize W with a fixed total energy per pulse

$$a_0^2 + \sum a_1^2 + b_1^2 = 1$$

and with boundary conditions $\varphi(1) = \varphi(-1) = 0$ we must obviously place all the energy in the first term, giving a cosine curve displaced to be tangent to the time axis.

$$\psi(t) = A [1 + \cos t]$$

$$A^2 \int_{-1}^1 (1 + \cos t)^2 dt = A^2 \left[t + 2 \sin t + \frac{t}{2} + \frac{1}{4} \sin 2t \right]_{-1}^1$$

$$= 2A^2 = 1$$

$$A = \frac{1}{\sqrt{2}}$$

$$\psi(t) = \frac{1}{\sqrt{2}} [1 + \cos t]$$

$$E = \frac{1}{2}$$

C. E. SHANNON

[24]

~~CONFIDENTIAL~~

(p7)

COVER SHEET FOR TECHNICAL MEMORANDA

RESEARCH DEPARTMENT

SUBJECT: A Mathematical Theory of Cryptography - Case 20878 (4)

ROUTING:

- 1 - HWB-HF-Case Files
- 2 - CASE FILES
- 3 - J. W. McRae
- 4 - L. Espenschied
- 5 - H. S. Black
- 6 - F. B. Llewellyn
- 7 - H. Nyquist
- 8 - B. M. Oliver
- 9 - R. K. Potter
- 10 - C. B. H. Feldman
- 11 - R. C. Mathes
- 12 - R. V. L. Hartley
- 13 - J. R. Pierce
- 14 - H. W. Bode
- 15 - R. L. Dietzold
- 16 - L. A. MacCall
- 17 - W. A. Shewhart
- 18 - S. A. Schelkunoff
- 19 - C. E. Shannon
- 20 - Dept. 1000 Files

MM- 45-110-92

DATE September 1, 1945

AUTHOR C. E. Shannon

INDEX NO. P 0.4

~~SECRET~~~~ABSTRACT~~

DOWNGRADED AT 3 YEAR INTERVALS
DECLASSIFIED AFTER 12 YEARS
DDO DAR 522A.14

ABSTRACT

A mathematical theory of secrecy systems is developed. Three main problems are considered. (1) A logical formulation of the problem and a study of the mathematical structure of secrecy systems. (2) The problem of "theoretical secrecy," i.e., can a system be solved given unlimited time and how much material must be intercepted to obtain a unique solution to cryptograms. A secrecy measure called the "equivocation" is defined and its properties developed. (3) The problem of "practical secrecy." How can systems be made difficult to solve, even though a solution is theoretically possible.

THIS DOCUMENT CONTAINS INFORMATION AFFECTING THE NATIONAL DEFENSE OF THE UNITED STATES WITHIN THE MEANING OF THE ESPIONAGE LAWS, TITLE 18 U.S.C. SECTIONS 793 AND 794. ITS TRANSMISSION OR THE REVELATION OF ITS CONTENTS IN ANY MANNER TO AN UNAUTHORIZED PERSON IS PROHIBITED BY LAW.

~~CONFIDENTIAL~~

MM-45-110-92

September 1, 1945

Index PO.4

MEMORANDUM FOR FILE

DOWNGRADED AT 3 YEAR INTERVALS
DECLASSIFIED AFTER 12 YEARS
DOD DIR 5200.10

Introduction and Summary

In the present paper a mathematical theory of cryptography and secrecy systems is developed. The entire approach is on a theoretical level and is intended to complement the treatment found in standard works on cryptography.* There, a detailed study is made of the many standard types of codes and ciphers, and of the ways of breaking them. We will be more concerned with the general mathematical structure and properties of secrecy systems.

The presentation is mathematical in character. We first define the pertinent terms abstractly and then develop our results as lemmas and theorems. Proofs which do not contribute to an understanding of the theorems have been placed in the appendix.

The mathematics required is drawn chiefly from probability theory and from abstract algebra. The reader is assumed to have some familiarity with these two fields. A knowledge of the elements of cryptography will also be helpful although not required.

The treatment is limited in certain ways. First, there are two general types of secrecy system; (1) concealment systems, including such methods as invisible ink, concealing a message in an innocent text, or in a fake covering cryptogram, or other methods in which the existence of the message is concealed from the enemy; (2) "true" secrecy systems where the meaning of the message is concealed by cipher, code, etc., although its existence is not hidden. We consider only the second type--concealment systems are more of a psychological than a mathematical problem. Secondly, the treatment is limited to the case of discrete information, where the information to be enciphered consists of a sequence of discrete symbols, each chosen from a finite set. These symbols may be letters in a

*See, for example, H.F.Gaines, "Elementary Cryptanalysis," or M. Givierge, "Cours de Cryptographie."

language, words of a language, amplitude levels of a "quantized" speech or video signal, etc., but the main emphasis and thinking has been concerned with the case of letters. A preliminary survey indicates that the methods and analysis can be generalized to study continuous cases, and to take into account the special characteristics of speech secrecy systems.

The paper is divided into three parts. The main results of these sections will now be briefly summarized. The first part deals with the basic mathematical structure of language and of secrecy systems. A language is considered for cryptographic purposes to be a stochastic process which produces a discrete sequence of symbols in accordance with some systems of probabilities. Associated with a language there is a certain parameter D which we call the redundancy of the language. D measures, in a sense, how much a text in the language can be reduced in length without losing any information. As a simple example, if each word in a text is repeated a reduction of 50 per cent is immediately possible. Further reductions may be possible due to the statistical structure of the language, the high frequencies of certain letters or words, etc. The redundancy is of considerable importance in the study of secrecy systems.

A secrecy system is defined abstractly as a set of transformations of one space (the set of possible messages) into a second space (the set of possible cryptograms). Each transformation of the set corresponds to enciphering with a particular key and the transformations are supposed reversible (non-singular) so that unique deciphering is possible when the key is known.

Each key and therefore each transformation is assumed to have an a priori probability associated with it--the probability of choosing that key. The set of messages or message space is also assumed to have a priori probabilities for the various messages, i.e., to be a probability or measure space.

In the usual cases the "messages" consist of sequences of "letters." In this case as noted above the message space is represented by a stochastic process which generates sequences of letters according to some probability structure.

These probabilities for various keys and messages are actually the enemy cryptanalyst's a priori probabilities for the choices in question, and represent his a priori knowledge of the situation. To use the system a key is first selected and sent to the receiving point. The choice of a key determines a particular transformation in the set forming the system. Then a message is selected and the particular transformation applied to this message to produce a cryptogram. This cryptogram is

transmitted to the receiving point by a channel that may be intercepted by the enemy. At the receiving end the inverse of the particular transformation is applied to the cryptogram to recover the original message.

If the enemy intercepts the cryptogram he can calculate from it the a posteriori probabilities of the various possible messages and keys which might have produced this cryptogram. This set of a posteriori probabilities constitutes his knowledge of the key and message after the interception.* The calculation of these a posteriori probabilities is the generalized problem of cryptanalysis.

As an example of these notions, in a simple substitution cipher with random key there are $26!$ transformations, corresponding to the $26!$ ways we can substitute for 26 different letters. These are all equally likely and each therefore has an a priori probability $1/26!$. If this is applied to "normal English" the cryptanalyst being assumed to have no knowledge of the message source other than that it is English, the a priori probabilities of various messages of N letters are merely their frequency in normal English text.

If the enemy intercepts N letters of cryptogram in this system his probabilities change. If N is large enough (say 50 letters) there is usually a single message of a posteriori probability nearly unity, while all others have a total probability nearly zero. Thus there is an essentially unique "solution" to the cryptogram. For N smaller (say $N = 15$) there will usually be many messages and keys of comparable probability, with no single one nearly unity. In this case there are multiple "solutions" to the cryptogram.

Considering a secrecy system to be a set of transformations of one space into another with definite probabilities associated with each transformation, there are two natural combining operations which produce a third system from two given systems. The first combining operation is called the product operation and corresponds to enciphering the message with the first system R and enciphering the resulting cryptogram with system S , the keys for R and S being chosen independently. This total operation is a secrecy system whose transformations consist of all the products (in the usual sense of products of transformations) of transformations in S with transformations in R . The probabilities are the products of the probabilities for the two transformations.

The second combining operation is "weighted addition"

$$T = pR + qS \quad p + q = 1$$

*"Knowledge" is thus identified with a set of propositions having associated probabilities. We are here at variance with the doctrine often assumed in philosophical studies which consider knowledge to be a set of propositions which are either true or false.

It corresponds to making a preliminary choice as to whether system R or S is to be used with probabilities p and q, respectively. When this is done R or S is used as originally defined.

It is shown that secrecy systems with these two combining operations form essentially a "linear associative algebra" with a unit element, an algebraic variety that has been extensively studied by mathematicians. Some of the properties of this algebra are developed.

Among the many possible secrecy systems there is one type with many special properties. This type we call a "pure" system. A system is pure if for any three transformations T_i , T_j , T_k in the set the product

$$T_i T_j^{-1} T_k$$

is also a transformation in the set, and all keys are equally likely. That is enciphering, deciphering, and enciphering with any three keys must be equivalent to enciphering with some key.

With a pure cipher it is shown that all keys are essentially equivalent--they all lead to the same set of a posteriori probabilities. Furthermore, when a given cryptogram is intercepted there is a set of messages that might have produced this cryptogram (a "residue class") and the a posteriori probabilities of messages in this class are proportional to the a priori probabilities. All the information the enemy has obtained by intercepting the cryptogram is a specification of the residue class. Many of the common ciphers are pure systems, including simple substitution with random key. In this case the residue class consists of all messages with the same pattern of letter repetitions as the intercepted cryptogram.

Two systems R and S are defined to be "similar" if there exists a fixed transformation A with an inverse, A^{-1} such that

$$R = AS.$$

If R and S are similar, a one-to-one correspondence between the resulting cryptograms can be set up leading to the same a posteriori probabilities. The two systems are cryptanalytically the same.

The second main part of the paper deals with the problem of "theoretical security." How secure is a system against cryptanalysis when the enemy has unlimited time and manpower available for the analysis or intercepted cryptograms?

~~CONFIDENTIAL~~

"Perfect Secrecy" is defined by requiring of a system that after a cryptogram is intercepted by the enemy the a posteriori probabilities of this cryptogram representing various messages be identically the same as the a priori probabilities of the same messages before the interception. It is shown that perfect secrecy is possible but requires, if the number of messages is finite, the same number of possible keys--if the message is thought of as being constantly generated at a given "rate" R , (to be defined later), key must be generated at the same or a greater rate.

If a secrecy system with a finite key is used, and N letters of cryptogram intercepted, there will be, for the enemy, a certain set of messages with certain probabilities, that this cryptogram could represent. As N increases the field usually narrows down until eventually there is a unique "solution" to the cryptogram--one message with probability essentially unity while all others are practically zero. A quantity $Q(N)$ is defined, called the equivocation, which measures in a statistical way how near the average cryptogram of N letters is to a unique solution; that is, how uncertain the enemy is of the original message after intercepting a cryptogram of N letters. Various properties of the equivocation are deduced--for example, the equivocation of the key never increases with increasing N . This quantity Q is a theoretical secrecy index--theoretical in that it allows the enemy unlimited time to analyse the cryptogram.

The function $Q(N)$ for a certain idealized type of cipher called the random cipher is determined. With certain corrections this function can be applied to many cases of practical interest. This gives a way of calculating approximately how much intercepted material is required to obtain a solution to a secrecy system. It appears from this analysis that with ordinary languages and the usual types of ciphers (not codes) this "unicity distance" is approximately $|K|/D$. Here $|K|$ is a number measuring the "size" of the key space. If all keys are a priori equally likely $|K|$ is the logarithm of the number of possible keys. D is the redundancy of the language and measures the excess information content of the language. In simple substitution with random key on English $|K|$ is $\log_{10} 26!$ or about 20 and D is about .7 for English. Thus unicity occurs at about 30 letters.

It is possible to construct secrecy systems with a finite key for certain "languages" in which the function $Q(N)$ does not approach zero as $N \rightarrow \infty$. In this case, no matter how much material is intercepted, the enemy still does not get a unique solution to the cipher but is left with many alternatives, all of reasonable probability. Such systems we call ideal systems. It is possible in any language to approximate such behavior--i.e., to make the approach to zero of $Q(N)$ recede

out to arbitrarily large N . However, such systems have a number of drawbacks, such as complexity and sensitivity to errors in transmission of the cryptogram.

The third part of the paper is concerned with "practical secrecy." Two systems with the same key size may both be uniquely solvable when N letters have been intercepted, but differ greatly in the amount of labor required to effect this solution. An analysis of the basic weaknesses of secrecy systems is made. This leads to methods for constructing systems which will require a large amount of work to solve. A certain incompatibility among the various desirable qualities of secrecy systems is discussed.

PART I

FOUNDATIONS AND ALGEBRAIC STRUCTURE OF SECRECY SYSTEMS

1. Choice, Information and Uncertainty

Suppose we have a set of possible events whose probabilities of occurrence are p_1, p_2, \dots, p_n . These probabilities are known, but that is all we know concerning which event will occur. Can we define a quantity which will measure in some sense how "uncertain" we are of the outcome? How much "choice" is involved in the selection of the event by the chance element that operates with these probabilities? We propose as a numerical measure of this rather vague notion the quantity

$$H = - \sum_{i=1}^n p_i \log p_i.$$

There are many reasons for this particular formula. Quantities of this kind appear continually in the present paper and in the study of the transmission of information.

To justify this definition we will state a number of properties that follow from it. These properties will not be proved here,* but are easily deduced from the definition. Properties of $H = - \sum p_i \log p_i$.

1. $H = 0$ if and only if all the p_i but one are zero, this one having the value unity. Thus only when we are certain of the outcome does H vanish.
2. For a given n , H is a maximum and equal to $\log n$ if and only if all the p_i are equal (i.e. $1/n$). This is also intuitively the most uncertain situation.
3. Suppose there are two events in question, with m possibilities for the first and n for the second. Let p_{ij} be the probability of the joint occurrence of i for the first and j for the second. The uncertainty of the joint event is

$$H = - \sum_{i,j} p_{ij} \log p_{ij}.$$

For given probabilities $p_i = \sum_j p_{ij}$ for the first and

* It is intended to develop these results in coherent fashion in a forthcoming memorandum on the transmission of information.

$q_j = \sum_i p_{ij}$ for the second, the quantity H is maximized if and only if the events are independent, i.e., $p_{ij} = p_i q_j$. This maximum value is the sum of the individual uncertainties

$$\begin{aligned} H &= H_1 + H_2 \\ &= -\sum p_i \log p_i - \sum q_j \log q_j. \end{aligned}$$

These facts can be generalized to any number of different events.

4. Suppose there are two chance events A and B as in 3, not necessarily independent. We define the mean conditional uncertainty of B , knowing A as

$$\bar{H}_A(B) = \sum_A p(A) H_A(B)$$

where $H_A(B)$ is the uncertainty of B when A has a definite value A . Thus $\bar{H}_A(B)$ is the average uncertainty of B for all different events A , weighted according to their different probabilities of occurrence. The uncertainty of the joint event is the sum of the uncertainty of the first and the mean conditional uncertainty of the second. In symbols

$$H(A, B) = H(A) + \bar{H}_A(B)$$

This is true whether or not there are any casual connections or correlations between the two events.

5. In the same situation the uncertainty of B is not greater than the joint uncertainty $H(A, B)$.

$$H(B) \leq H(A, B)$$

The equality holds if and only if every B (of probability greater than zero) is consistent with only one A . That is, if A is uniquely determined by B .

6. From properties 3 and 4 we have

$$\begin{aligned} H(A) + H(B) &\geq H(A, B) \\ H(B) &\geq H(A, B) - H(A) \\ &= H(A) + \bar{H}_A(B) - H(A) \\ H(B) &\geq \bar{H}_A(B) \end{aligned}$$

Thus the uncertainty of B is not greater than its average value when we know A. Additional information never increases average uncertainty. The equality holds if and only if A and B are independent.

7. Suppose we have a set of probabilities p_1, p_2, \dots, p_n . Any change toward equalization of these (supposing them unequal) increases H. Thus if $p_1 < p_2$ and we increase p_1 , decreasing p_2 an equal amount (to keep the sum $\sum p_i$ constant at unity) so that p_1 and p_2 are more nearly equal, then H increases. More generally if we perform any "averaging" operation on the p_i of the form

$$p_i' = \sum a_{ij} p_j$$

where $\sum_i a_{ij} = 1$ and all $a_{ij} \geq 0$, then H increases (except in the special case where this transformation amounts to no more than a permutation of the p_j with H of course remaining the same).

8. H measures in a certain sense how much "information is generated" when the choice is made. Suppose such a chance event occurs and we wish to describe which of the n possible events took place. The average amount of paper required to write it down in a properly chosen notation is in the cases of interest to us, about proportional to H. Thus there might be $10^{30} + 10^{50}$ possible events, with 10^{30} of them having a probability $\frac{1}{2} 10^{-30}$ and 10^{50} a probability of $\frac{1}{2} 10^{-50}$. We could set up a notational system to describe which event occurs as follows. We number the events from 1 up to $10^{30} + 10^{50}$ and when one occurs write down the corresponding number. The average amount of paper required will be proportional to the average number of digits we need. This will be nearly 30 if the event is in the first group of 10^{30} and about 50 if in the second group. Thus the average number of digits is about 40. We also have

$$H = -10^{30} \frac{1}{2} 10^{-30} \log \frac{1}{2} 10^{-30} - 10^{50} \frac{1}{2} 10^{-50} \log \frac{1}{2} 10^{-50} \\ = 40$$

9. Although the last result is only approximately true when the number of choices is finite it becomes exactly true when an unlimited sequence of choices is made. Thus if a sequence of N independent choices is made each choice being from n possibilities with probabilities p_1, p_2, \dots, p_n then the total amount of information generated is

$$H = -N \sum p_i \log p_i$$

If N is sufficiently large, the expected number of digits required to register the particular choice made is arbitrarily close to H , providing the correspondence between sequences of digits and sets of choices is correctly made. If incorrectly made it will be greater than H . Moreover if N is sufficiently large the probability of needing more than H digits is very small.

10. It can be shown that if we require certain reasonable properties of a measure of choice or uncertainty then formula $-\sum p_i \log p_i$ necessarily follows. These requirements and the proof of this statement are given in Appendix I. The chief property is that the measure be a sense additive--if a choice be decomposed into a series of choices the total choice is the sum (properly weighted) of the individual choices.

11. Finally we note that quantities of the type $\sum p_i \log p_i$ have appeared previously as measures of randomness, particularly in statistical mechanics. Indeed the H in Boltzmann's theorem is defined in this way, p_i being the probability of a system being in cell i of its phase space. Most of the entropy formulas contain terms of this type.

The base which is used in taking logarithms in the formulas amounts to a choice of the unit of measure. If the base is 2 we will call the resulting units "digits;" if the base is 10 the units will be called "alternatives." One digit is about 3.32 alternatives. A choice from 1000 equally likely possibilities is 3 digits or about 10 alternatives.

2. Language as a Stochastic Process

A natural language, such as English, can be studied from many points of view--lexicography, syntax, semantics, history, aesthetics, etc. The only properties of a language of interest in cryptography are statistical properties. What are the frequencies of the various letters, of different digrams (pairs of letters), trigrams, words, phrases, etc.? What is

the probability that a given word occurs in a certain message. The "meaning" of a message has significance only in its influence on these probabilities. For our purposes all other properties of language can be omitted. We consider a language therefore, to be a stochastic (i.e. a statistical) process which generates a sequence of symbols according to some system of probabilities. The symbols will be the letters of the language together with punctuation, spaces, etc., if these occur.

Conversely any stochastic process which produces a discrete sequence of symbols will be said to be a language. This will include such cases as:

1. Natural written languages such as English, German, Chinese
2. Continuous information sources that have been rendered discrete by some quantizing process. For example, the quantized speech from a PCM transmitter, or a quantized television signal.
3. "Artificial" languages, where we merely define abstractly a stochastic process which generates a sequence of symbols. The following are examples of artificial languages.

- (A) Suppose we have 5 letters A, B, C, D, E which are chosen each with probability .2, successive choices being independent. This would lead to a sequence in which the following is a typical example.

B D C B C E C C C A D C B D D A A E C E E A
A B B D A E E C A C E E B A E E C B C E A D

This was constructed with the use of a table of random numbers.*

- (B) Using the same 5 letters let the probabilities be .4, .1, .2, .2, .1 respectively, with successive choices independent. A typical "text" in this language is then:

A A A C D C B D C E A A D A D A C E D A
E A D C A B E D A D D C E C A A A A A D

- (C) A more complicated structure is obtained if successive letters are not chosen independently but their probabilities depend on preceding letters. In the simple

* Kendall and Smith, "Tables of Random Sampling Numbers," Cambridge, 1939.

case of this type a choice depends only on the preceding letter and not on ones before that. The statistical structure can then be described by a set of transition probabilities $p_i(j)$, the probability that letter i is followed by letter j . The indices i and j range over all the letters in the language. A second equivalent way of specifying the structure is to give the digram probabilities $p(i,j)$, the relative frequency of the digram ij in the language. The letter frequencies $p(i)$, the probability of letter i , the transition probabilities $p_i(j)$ and the digram probabilities $p(i,j)$ are related by the following formulas.

$$p(i) = \sum_j p(i,j) = \sum_j p(j,i) = \sum_j p(j)p_j(i)$$

$$p(i,j) = p(i) p_i(j)$$

$$\sum_j p_i(j) = \sum_i p(i) = \sum_{i,j} p(i,j) = 1$$

As a specific example suppose there are three letters A, B, C with the probability tables:

$p_i(j)$	j			$p(i)$		$p(i,j)$	j	
	A	B	C				A	B
A	0	.8	.2	$\frac{9}{27}$	A	0	$\frac{4}{15}$	
B	.5	.5	0	$\frac{16}{27}$	B	$\frac{8}{27}$	$\frac{8}{27}$	
C	.5	.4	.1	$\frac{2}{27}$	C	$\frac{1}{27}$	$\frac{4}{135}$	

A typical text in this language is the following.

ABBABABABABABABBBBABBBBABB
 ABABABABBBBACACABBABBBBABB
 ABACBBBABA

The next increase in complexity would involve trigram frequencies but no more. The choice of a letter would depend on the preceding two letters but not on the text before that point. A set of trigram frequencies

$p(i,j,k)$ or equivalently a set of transition probabilities $p_{ij}(k)$ would be required. Continuing in this way one obtains successively more complicated stochastic processes. In the general n -gram case a set of n -gram probabilities $p(i_1, i_2, \dots, i_n)$ or of transition probabilities $p_{i_1, i_2, \dots, i_{n-1}}$ is required to specify the statistical structure.

- (D) Stochastic processes can also be defined which produce a text consisting of a sequence of "words." Suppose there are 5 letters A, B, C, D, E and 16 "words" in the language with associated probabilities:

.10 A	.16 BEBE	.11 CABED	.04 DEB
.04 ADEB	.04 BED	.05 CEED	.15 DEED
.05 ADEE	.02 BEED	.08 DAB	.01 EAB
.01 BADD	.05 CA	.04 DAD	.05 EE

Suppose successive "words" are chosen independently and are separated by a space. A typical message might be:

DAB EE A BEBE DEED DEB ADEE ADEE EE DEB BEBE BEBE
 BEBE ADEE BED DEED DEED CEED ADEE A DEED DEED BEBE
 CABED BEBE BED DAB DEED ADEB

If all the words are of finite length this process is equivalent to one of the preceding type, but the description may be simpler in terms of the word structure and probabilities. We may also generalize here and introduce transition probabilities between words, etc.

These artificial languages are useful in constructing simple problems and examples to illustrate various possibilities. We can also approximate to a natural language by means of a series of simple artificial languages. The zero order approximation is obtained by choosing all letters with the same probability and independently. The first order approximation is obtained by choosing successive letters independently but each letter having the same probability that it does in the natural language. Thus in the first order approximation to English is chosen with probability .12 (its frequency in normal English) and W with probability .02, but there is no influence between adjacent letters and no tendency to form the preferred digrams such as TH, ED, etc. In the second order approximation digram structure is introduced. After a letter is chosen, the next

one is chosen in accordance with the frequencies with which the various letters follow the first one. This requires a table of digram frequencies $p_1(j)$, the frequency with which letter j follows letter i . In the third order approximation trigram structure is introduced. Each letter is chosen with probabilities which depend on the preceding two letters.

3. The Series of Approximations to English

To give a visual idea of how this series of processes approaches a language, typical sequences in the approximation to English have been constructed and are given below. In all cases we have assumed a 27 symbol "alphabet," the 26 letters and a space.

1. Zero order approximation (symbols independent and equally probable).

XFOML RXKHRJFFJUF ZLPWCFWKCXYJ FFJEYVKCQSGXYD
QPAAMKBZAACIBZLHJQD

2. First order approximation (symbols independent but with frequencies of English text).

OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHT
OObTTVA NAH BRL

3. Second order approximation (digram structure as in English).

ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D
ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE
SEACE CTISBE

4. Third order approximation (trigram structure as in English).

IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOL
OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE

5. 1st Order Word Approximation. Rather than continue with tetragram, ..., n-gram structure it is easier and better to jump at this point to word units. Here words are chosen independently but with their appropriate frequencies.

REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN
DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO
EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE
THESE.

6. 2nd Order Word Approximation. The word transition probabilities are correct but no further structure is included.

THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED

The resemblance to ordinary English text increase quite noticeably at each of the above steps. Note that the samples have reasonably good structure out to about twice the range that is taken into account in their construction. In (3) the statistical process insures reasonable text for letter sequence, but four-letter sequences from the sample usually be fitted into good sentences. In (6) sequences of or more words can easily be placed in sentences without un- or strained constructions. The particular sequence of ten words "attack on an English writer that the character of th is not at all unreasonable.

The first two samples were constructed by the use of a book of random numbers in conjunction for (2) with a table of letter frequencies. This method might have been continued for (3), (4), and (5), since digram, trigram, and word frequency tables are available, but a simpler equivalent method was used. To construct (3) for example one opens a book at random and selects a letter at random on the page. This letter is recorded. The book is then opened to another page and one reads until this letter is encountered. The succeeding letter is then recorded. Turning to another page this second letter is searched for and the succeeding letter recorded, etc. A similar process was used for (4), (5), and (6). It would be interesting if further approximations could be constructed, but the labor involved becomes enormous at the next stage.

The stochastic process 6 is already sufficiently close to English for many cryptographic purposes since most cryptanalysis is based on "local" structure of not more than two or three words in length.

4. Graphical Representation of a Markoff Process

Stochastic processes of the type described above are known mathematically as discrete Markoff processes and have been extensively studied in the literature.* The general case

* For a detailed treatment see M. Frechet, "Methods des fonctions arbitraires. Theorie des évenements en chaîne dans le cas d'un nombre fini d'états possibles." Paris, Gauthier-Villars, 1938.

can be described as follows. There exist a finite number c possible "states" of a system; S_1, S_2, \dots, S_n . In addition there is a set of transition probabilities; $q_i(j)$ the probability that if the system is in state S_i it will next go to state S_j . To make this Markoff process into a language generator we need only assume that a letter is produced for each transition from one state to another. The states will correspond to the "residue of influence" from preceding letters.

The situation can be represented graphically as shown in Figs. 1, 2, 3 and 4. The "states" are the junction points in the graph and the probabilities and letters produced for transition are given beside the corresponding line. Fig. 1 for the example B in Section 2, while Fig. 2 corresponds to example C. In Fig. 1 there is only one state since successive letters are independent. In Fig. 2 there are as many states as letters. If a trigram example were constructed there would be at most n^2 states corresponding to the possible pairs of letters preceding the one being chosen. Figs. 3 and 4 show graphs for the case of word structure in example D. In Fig. 3 S corresponds to the "space" symbol. In Fig. 3 each word has a separate chain of branches from the left to the right junction point, while in Fig. 4 the branches have been combined, simplifying the graph.

5. Pure and Mixed Languages

As we have indicated above a "language" for our purposes can be considered to be generated by a Markoff process. Among the possible discrete Markoff processes there is a group with special properties of significance in cryptographic work. This special class consists of the "ergodic" processes and we shall call the corresponding languages "pure languages." Although a rigorous definition of an ergodic process is somewhat involved, the general idea is simple. In an ergodic process every sequence produced by the process is the same in statistical properties. Thus the letter frequencies, digram frequencies, etc., obtained from particular sequences will, as the lengths of the sequences increase, approach definite limits independent of the particular sequence. Actually this is not true of every sequence but the set for which it is false has probability zero. Roughly the ergodic property means statistical homogeneity.

All the examples of artificial languages given above are pure, the corresponding Markoff process being ergodic. This property is related to the structure of the corresponding graph. If the graph has two properties the language it generates will be pure. These properties are:

1. The graph cannot be divided into two parts A and B such that it is impossible to go from junction points in part A to junction points in part B along lines of the graph in the direction of arrows and also impossible to go from nodes in part B to nodes in part A.
2. A closed series of lines in the graph with all arrows on the lines pointing in the same orientation will be called a "circuit." The "length" of a circuit is the number of lines in it. Thus in Fig. 4 the series BEB is a circuit of length 4. The second property required is that the greatest common divisor of the lengths of all circuits in the graph be one.

If the first condition is satisfied but the second one violated by having the greatest common divisor equal to $d > 1$, the sequences have a certain type of periodic structure. The various sequences fall into d different classes which are statistically the same apart from a shift of the origin (i.e., which letter in the sequence is called letter 1). By a shift of from 0 up to $d - 1$ any sequence can be made statistically equivalent to any other. A simple example with $d = 2$ is the following. There are three possible letters a, b, c. Letter a is followed with either b or c with probabilities $\frac{1}{3}$ and $\frac{2}{3}$ respectively. Either b or c is always followed by letter a. Thus a typical sequence is

a b a c a c a c a b a c a b a b a c a c . .

This type of situation is not of much importance for our work.

If the first condition is violated the graph may be "separated" into a set of subgraphs each of which satisfies the first condition. We will assume that the second condition is also satisfied for each subgraph. We have in this case what may be called a "mixed" language made up of a number of pure components. The components correspond to the various subgraphs. If L_1, L_2, L_3, \dots are the component languages we may write

$$L = p_1 L_1 + p_2 L_2 + p_3 L_3 + \dots$$

where p_1 is the a priori probability of the component language L_1 .

Physically the situation represented is this. There are several different languages L_1, L_2, L_3, \dots which are of homogeneous statistical structure (i.e., they are pure languages). We do not know a priori which is to be used, but once the sequence starts in a given pure component L_i it con-

indefinitely according to the statistical structure of that component. We do have, however, a set of a priori probabilities for the various components, p_1, p_2, \dots .

As an example one may take two of the artificial languages defined above and assume $p_1 = .2$ and $p_2 = .8$. A sequence from the mixed language

$$L = .2 L_1 + .8 L_2$$

would be obtained by choosing first L_1 or L_2 with probabilities .2 and .8 and after this choice generating a sequence from whichever was chosen.

A natural language, such as English or German, is not, of course, pure. Different kinds of text, literary, newspaper, technical or military, display consistently different types of structure. These differences are small, however, in comparison with the differences between different natural languages. If only local structure--letter, digram and trigram frequencies, for instance--is of much importance, it is reasonable to consider "normal English" to be nearly pure.

6. Information Rate and Redundancy of a Language

Suppose we have a pure language L produced by a given Markoff process. Associated with the language there are certain parameters which are of significance in questions of transforming the language and in cryptography. The most important of these is what we will call the "information rate" R for the language. It measures the rate at which the Markoff process "generates information," as determined by the measurement of the amount of choice available on the average per letter of text that is produced. In Section 1 we defined the amount of choice when there are various possibilities with probabilities p_1, p_2, \dots, p_n as

$$H = - \sum p_i \log p_i .$$

In a Markoff process with a number of different "states" there will be a choice value H_i for each of these states and a probability of being in each of the states (or a frequency with which this state occurs). If this relative frequency for state i is P_i , the average amount of choice is

$$R = \sum P_i H_i$$

summed over all the states. This is the definition of the

~~CONFIDENTIAL~~

information rate for the language. If $p_i(j)$ is the probability of producing letter j when in state i we have

$$H_i = \sum_j p_i(j) \log p_i(j)$$

the sum being over all the letters in the language. Thus

$$R = \sum_i \sum_j P_i p_i(j) \log p_i(j)$$

The information rate R has the units of alternatives (or digits) per letter since it measures the average amount of choice per letter of text that is produced.

A second parameter of importance is the "maximum rate" R_0 for the source. This is defined simply as the logarithm of the number of different letters in the language. R_0 is also measured in alternatives or digits per letter. If successive letters are chosen independently and each letter is equally likely $R_0 = R$. Otherwise we have $R < R_0$.

R and R_0 are actually two limiting cases of information rates for the language. R_0 may be said to be the rate when no statistical structure is taken into consideration and R is the rate when all the structure is taken into account. Between these there is an infinite series of rates $R_1, R_2, \dots, R_n, \dots$ which take some of the statistical structure into account. R_1 takes the letter frequencies into account and is defined by

$$R_1 = \sum p(i) \log p(i)$$

where $p(i)$ is the probability of letter i . R_2 takes digram structure into account and is defined by

$$R_2 = \sum p(i) \sum p_i(j) \log p_i(j)$$

where the $p(i)$ are letter probabilities and $p_i(j)$ the transition probabilities, i.e., the probability of letter i being followed by letter j . In general we define

$$R_n = \sum p(i_1, i_2, \dots, i_{n-1}) \log p_{i_1 i_2 \dots i_{n-1}}(i_n)$$

where the sum is on all indices i_1, \dots, i_n and $p_{i_1} \dots i_{n-1}$ is the probability of $(n-1)$ gram $i_1 \dots i_{n-1}$ with $p_{i_1} \dots i_{n-1}(i_n)$ the probability of this $n-1$ gram being followed by letter i_n . R_n may be called the n -gram information rate for the language. It can be shown that

$$R_0 \geq R_1 \geq R_2 \geq \dots \geq R_n \geq \dots R_\infty = R$$

These rates determine how much a language can be "compressed" in length by a suitable encoding process. A language with maximum rate R_0 and rate R can be transformed in such a way that a sequence of letters N letters long is transformed into a sequence of letters only N' letters long where

$$N' R_0 = N R$$

(This is approximate and only exactly true in the limit as $N \rightarrow \infty$.) Thus the information is "compressed" in the ratio

$$\frac{R}{R_0}$$

This is the greatest compression ratio possible. It makes use of all the statistical structure of the language. If only n -gram structure is made use of, a compression ratio

$$\frac{R_n}{R_0}$$

is the best possible.

The compression obtained in this way is only a statistical gain. Some infrequent sequences are encoded into much longer sequences while the more probable ones go into shorter sequences so that on the average the length is decreased. It is the type of compression obtained in telegraphy by using the shortest telegraph symbol, a single dot, for the most frequent letter E, while uncommon letters Q, Z, etc., are encoded into longer telegraph symbols. An average reduction in time of transmission is obtained but there are possible sequences, e.g., Q Q Q : : :, which require much longer.

Performing a transformation on a language L which compresses as much as possible will be called reducing L to a "normal" form. When this has been done it can be shown that all letters in the output are equally likely and independent. Actually to realize this transformation would usually

require an infinitely complex machine, but we can always approximate it as closely as desired with a machine of finite complexity.

The quantity

$$D = R_0 - R$$

will be called the redundancy rate of the language. It means the excess information that is sent if sequences in the language are transmitted in their original form (without compression reduction to normal form). Correspondingly there is a whole series of redundancy rates:

$$D_0 = R_0 - R_0 = 0$$

$$D_1 = R_0 - R_1$$

$$D_2 = R_1 - R_2$$

.

.

$$D_n = R_0 - R_n$$

.

.

$$D = R_0 - R$$

D_n is the redundancy rate due to n-gram structure in the language.

The redundancy D can also be said to measure the amount of statistical structure in the language. If the sequence is purely random $D = 0$ while at the other extreme if each letter is completely determined by preceding letters with no freedom of choice, D has its maximum possible value R_0 . It is sometimes convenient to use the "relative" redundancy D/R_0 which must lie between 0 and 100%.

If we have a source of rate R , maximum rate R_0 (both in digits per letter) and consider the possible sequences of letters these fall into two groups for N large. One group of "high probability" sequences contains about

$$10^{RN}$$

sequences (where we have assumed R measured in digits per letter). All of these have substantially the same logarithmic probability.

The remainder of the total of $10^{R_0 N}$ possible sequences are of very small probability. In fact their total probability approaches zero as N increases. The logarithm of the probability of an individual sequence in the high probability group is thus about $-RN$. In a precise statement of these results we must allow a certain fuzziness in R , i.e., replace R by $R \pm \epsilon$ where $\epsilon \rightarrow 0$ as $N \rightarrow \infty$.

Reduction of a language to normal form is performed by properly matching the probabilities of sequences to the length of the corresponding sequences in the normal form. The "high probability" sequences are translated into short sequences and the remainder into longer sequences.

An example will clarify the results we have given. Let the language contain 4 letters A, B, C, D. In a sequence successive letters are chosen independently, the four letters having probabilities $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}$, respectively. We have

$$R_0 = \log_2 4 = 2 \text{ alternatives/letter}$$

and

$$R_1 = R_2 = R_3 = \dots = R = - \left(\frac{1}{2} \log \frac{1}{2} + \frac{1}{4} \log \frac{1}{4} + \frac{2}{8} \log \frac{1}{8} \right)$$

$$= \frac{1}{2} + \frac{1}{2} + \frac{6}{8} = \frac{7}{4} \text{ alternatives/letter}$$

By a suitable transformation the average length of sequences can be reduced by the factor $\frac{7}{4}/2 = 7/8$. A transformation to do it is the following. First we translate into a sequence of binary digits (0 or 1) by the following table

A	0
B	10
C	110
D	111

After this pairs of the binary digits are translated into the original alphabet as follows

00	A'
01	B'
10	C'
11	D'

For a typical sequence this works out as shown below:

Translation into binary digits:

A	B	C	A	B	A	C	B	B	D	A	A	D	A	D	A
0	10	110	0	10	0	110	10	10	111	0	0	111	0	111	0

Regrouping and translation back into letters:

01	01	10	01	00	11	01	01	01	11	00	11	10	11	10
B'	B'	C'	B'	A'	D'	B'	B'	B'	D'	A'	D'	C'	D'	C'

In this case there are 16 letters in the original and 15 in final text. Thus due to the small redundancy and the short of the text only part of the saving is evident. In a long however the full reduction of $\frac{1}{8}$ would appear. This may be verified directly in this case. In a long text of N letter each letter will appear with about its appropriate frequency. Thus the number of binary digits will be about

$$N[\frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \frac{1}{8} \cdot 3] = \frac{7}{4} N$$

since each A gives one binary digit, each B gives two, etc. number of letters in the final text is half this since each pair of binary digits goes into one letter. Thus the reduction is by a factor $\frac{7}{8}$.

It is also easy to see in this case that the binary digits are equally likely and independent, and from this the final text letters are also.

This situation is more complicated for mixed long and we shall not enter into it here. We may note, however, that if

$$L = p_1 L_1 + p_2 L_2 + \dots + p_n L_n$$

where L_1 is pure with rate $R^{(1)}$, then the long sequences of fall into $(n+1)$ groups. The first n groups correspond to the pure components. Those in group 1 number about

$$10^{R^{(1)}N}$$

and have logarithmic probability about

$$- R^{(1)}N$$

The last group contains all other sequences and has a small total probability.

7. Redundancy Characteristic of a Language

The form of the curve $D(N)$ as a function of N may be called the redundancy characteristic of the language. In a rough way it describes the way in which the redundancy appears. In Fig. 5 several types of characteristics are shown, all with the same final redundancy. The way in which this approach is of importance in cryptography. For languages which reach final redundancy at one or two letters (Curves 1 and 2) one type of cipher (ideal ciphers) can be used. For those which remain near zero out to fairly large N (like Curve 5) another type is appropriate. Natural languages are apt to show a characteristic more like 3, and this makes them difficult to encipher with security by simple means.

Examples:

1. A language in which successive letters are independent but with different probabilities has a characteristic Type 1.
2. Consider a language constructed as follows. First select 26^8 different sequences of letters, each 16 letters long, from the 26^{16} possible sequences of this length. This should be a random selection. The 16-letter sequences chosen are the "words" of the language. Messages are random sequences of these "words." Such a language has a characteristic like the Curve 5.
3. A language with digram structure only, such as Example in Section 2 above, has a characteristic of the Type shown in Fig. 5, reaching its final value at $N = 2$.
4. English has the characteristic 3 in Fig. 5.

The redundancy characteristic describes how the structure in the language is spread out. If the structure is localized, the curve rises rapidly to its final value. If there are long range influences the asymptotic value is approached more slowly. If the structure is "locally random" the curve will remain near zero for small N .

8. Secrecy Systems

Before we can apply any mathematical analysis to secrecy systems, it is necessary to idealize the situation suitably, and to define in a mathematically acceptable way what we shall mean by a secrecy system. A "schematic" diagram of a general secrecy system is shown in Fig. 6. At the transmitting end there are two information sources--a message source and a key source. The key source produces a particular key from among those which are possible in the system. This key is transmitted by some means, supposedly not interceptible, e.g. by messenger, to the receiving end. The message source produces a message (the "clear") which is enciphered, and the resulting cryptogram sent to the receiving end by a possibly interceptible means, for example radio. At the receiving end the cryptogram and key are combined in the decipherer to recover the message.

Evidently the encipherer performs a functional operation. If M is the message, K the key, and E the enciphered message, or cryptogram, we have

$$E = f(M, K)$$

i.e. E is a function of M and K . We prefer to think of this, however, not as a function of two variables but as a (one parameter) family of operations or transformations, and we write it

$$E = T_i M.$$

The transformation T_i applied to message M produces cryptogram E . The index i corresponds to the particular key being used. If there are m possible keys there will be m transformations in the family T_1, T_2, \dots, T_m .

At the receiving end it must be possible to recover M , knowing E and K . Thus the transformations in the family must have unique inverses

$$M = T_i^{-1} E$$

at any rate this inverse must exist uniquely for every E which can be obtained from an M with key i .

The key source can be thought of as a "probability machine," something which chooses from the possible keys according to a system of probabilities. Mathematically then, the keys (or the parameter of the family of transformations) belong

~~CONFIDENTIAL~~

to a probability or measure space. Hence we arrive at the definition:

A secrecy system is a family of uniquely reversible transformations T_i of a message space \mathcal{M} into a cryptogram space \mathcal{C} , the parameter i belonging to a probability space \mathcal{K} . Conversely any set of entities of this type will be called a "secrecy system."

The system can be visualized mechanically as a machine with one or more controls on it. A sequence of letters, the message, is fed into the input of the machine and a second series emerges at the output. The particular setting of the controls corresponds to the particular key being used. Some method must be prescribed for choosing the key from all the possible ones.

To make the problem mathematically tractable we shall assume that the enemy knows the system being used. That is, he knows the family of transformations T_i , and the probabilities of choosing various keys.

One might object to this as being unrealistic, in that the cryptanalyst often does not know what system was used or the probabilities of various keys. There are two answers to this objection.

1. The assumption is actually the one ordinarily used in cryptographic studies. It is pessimistic and hence safe, but in the long run realistic (particularly in military work), since one must expect his system to be found out eventually through espionage, captured equipment, prisoners, etc. Thus, even when an entirely new system is devised, so that the enemy cannot assign any a priori probability to it without discovering it himself, one must still live with the expectation of his eventual knowledge.
2. The restriction is much weaker than appears at first, due to our broad definition of what constitutes the system. Suppose a cryptographer intercepts a message and does not know whether a substitution, transposition, or Vigenere type cipher was used. He can consider this as being enciphered by a system in which part of the key is the specification of which of these types was used, the next part being the particular key for that type. These three different possibilities are assigned probabilities according to his best guesses of the a priori probabilities of the encipherer using the respective types of cipher.

A second possible objection to our definition of secrecy systems is that no account is taken of the common practice of inserting nulls in a message and the use of multiple substitutes. Thus there is not a unique $E = T, M, k$, actually the encipherer can choose at will among a number of different E 's for the same message and key. This situation could be handled, but would only add complexity at the pre stage, without altering any of the basic results. To define the more general secrecy system, one would add a second parameter to the transformations T_i , which corresponds to the various choices of cryptograms corresponding to a given message and key. It is possible, but not always desirable, to consider this second parameter as part of the key, since it does not need to be transmitted to the receiving point.

We also assume that the enemy is in possession of a measure in the space Ω_M , the a priori probabilities of various messages. The same objection and essentially the same answer might be given to this assumption as to his knowledge of the transformations T_i . This measure, however, we do not consider as part of the secrecy system for reasons which will appear later. The secrecy system whose transformations are T_i will be denoted by T and this concept includes the space Ω_M on which T operates (without its measure), the transformation T_i and the spaces Ω_K and Ω_E , the former with its probability measure.

If the messages are produced by a Markoff process of the type described previously, the probabilities of various messages are determined by the structure of the Markoff process. For the present, however, we wish to take a more general view of the situation and regard the messages as merely an abstract set of entities with associated probabilities, not necessarily composed of a sequence of letters and not necessarily produced by a Markoff process.

It should be emphasized that throughout the paper a secrecy system means not one but a set of many transformations. After the key is chosen only one of these transformations is used and we might be led to define a secrecy system as a single transformation on a language.* The enemy, however, does not know what key was chosen and the "might have been" keys are as important for him as the actual one. Indeed it is only the existence of these other possibilities that gives the system its security.

*A. A. Albert in a paper presented at a Manhattan, Kansas, meeting of the American Mathematical Society (Nov. 22, 1944) entitled "Some Mathematical Aspects of Cryptography," has defined a ciphering system in this way. With this limited definition about all one can do is to describe and classify from the mathematical point of view various types of transformations.

any secrecy. Since the secrecy is our primary interest, we are forced to this rather elaborate concept of a secrecy system. This type of situation where possibilities are as important as actualities is almost the rule in games of strategy. The course of a chess game is largely controlled by threats which are not carried out. See also the "virtual existence" of unrealized imputations in von Neumann's theory of games.

There are a number of difficult epistemological questions connected with the theory of secrecy, or in fact with any theory which involves questions of probability (particularly a priori probabilities, Bayes' theorem, etc) when applied to a physical situation. Treated abstractly probability theory can be put on a rigorous logical basis with the modern measure theory approach.* As applied to reality, however, especially when "subjective" probabilities and unpredictable experiments are concerned, there are many questions of logical validity. For example in the approach to secrecy made here, a priori probabilities of various k are assumed known by the enemy cryptographer--how can one determine operationally if his estimates are correct, on basis of his knowledge of the situation?

It may happen that the keys are chosen by the cipherer according to one system of probabilities, i.e. a measure in the key space Ω_K and that the enemy cryptanalyst estimates a second different system of probabilities Ω'_K in this space which are entirely reasonable in the light of his knowledge of the situation-- which is correct? I believe that both are correct. The calculation based on Ω_K leads to the solution when the enemy knows just how the keys are chosen and the solution based on Ω'_K leads to solutions which are correct for a situation agreeing with the enemy's knowledge of the actual situation. It appears intuitively that the enemy's lack of knowledge can only do him harm, and probably this can be proved, but this question has not been investigated. In fact, we assume only one measure Ω_K in the key space. Similar remarks may be made regarding measure in the message space Ω_M .

*See J. L. Doob, "Probability as Measure," Annals of Math Stat.; v. 12, 1941, pp. 206-214.
A. Kolmogoroff, "Grundbegriffe der Wahrscheinlichkeitsrechnung," Ergebnisse der Mathematik, v. 2, No. 3 (Berlin 1933).

Actually in practical situations, only extreme errors in a priori probabilities of keys and messages can much error in the important parameters. This is because the exponential behavior of the number of messages, etc., and the logarithmic measures employed.

With regard to the application of the mathematical theory of probability to physical situations there are two main theories or ways of setting up the correspondence. The frequency theory. Probability is correlated with the frequency of an event. This is the correspondence used by the practicing statistician, in principle by the physicist etc. (2) The degree of belief approach. Probability is a subjective phenomena and measures one's degree of belief in the occurrence of an event. This approach is seen often in the work of historians, judges, and in everyday life. Although this latter approach has often been attacked as meaningless we cannot agree with this opinion. In the first place the intuitive approach can be given a rigorous mathematical foundation. This has been done in a very elegant way by B. O. Koopman.* Essentially one need only assume that a man is capable of making probability judgments (Event A is more probable than event B or they are equiprobable) and that his judgments be self consistent (e.g. if he judges A more probable than B and B more probable than C he should judge A more probable than C). One can even establish numerical values by the use of a "standard gauge," for example a roulette wheel and thus relate the subjective and the frequency probability. In the second place, on pragmatic grounds one can hardly deny the subjective applications, since almost all of our everyday decisions are based on this sort of probability judgment. Cryptographic work involves both types of applications. In the use of frequency tables, significance tests etc., the cryptanalyst is following the frequency approach. In the "intuitive" methods of cryptanalysis (probable words etc) the degree of belief approach is more in evidence.

We may remark that a single operation on a language which is reversible forms a degenerate type of system under our definition--a system with only one key and unit probability. Such a system has no secrecy--the cryptanalyst finds the message by applying the inverse of this transformation, the only one in the system, to the intercepted cryptogram. The decipherer and cryptanalyst in this case

*B. O. Koopman, "The Axioms and Algebra of Intuitive Probability," Annals of Mathematics, v.41, no.2, 1940, p.269. "Intuitive Probabilities and Sequences," v.42, no.1, 1941, p.169.

possess the same information. In general, the only difference between the decipherer's knowledge and the enemy cryptanal knowledge is that the decipherer knows the particular key used, while the cryptanalyst only knows the a priori probabilities of the various keys in the set. The process of deciphering is that of applying the inverse of the particular transformation used in enciphering to the cryptogram. The process of cryptanalysis is that of attempting to determine the message (or the particular key) given only the cryptogram and the a priori probabilities of various keys and messages.

A system will be called "closed" if any possible cryptogram can be deciphered with any possible key. This means that the inverse transformations T^{-1}_i are all defined for every element in the cryptogram space.

We shall use the notation $|M|$ for the "size" of message space:

$$|M| = -\sum P(M) \log P(M)$$

where $P(M)$ is the probability of message M and the sum is over all messages of just N letters. Thus $|M|$ is a function of N and measures the amount of "choice" in the selection of an N -letter message. For large N , $|M|$ is approximately RN . Similarly $|K|$ is the size of the key space

$$|K| = -\sum P(K) \log P(K)$$

the sum being over all keys.

9. Representation of Systems

A secrecy system can be represented in various ways. One which is convenient for illustrative purposes is a line diagram, as in Figs. 7, 10, 11. The possible messages are represented by points at the left and the possible cryptograms by points at the right. If a certain key, say key 1, transforms message M_2 into cryptogram E_4 , then M_2 and E_4 are connected by a line labeled 1, etc. From each possible message there must be exactly one line emerging for each different

A second representation is by means of a rectangular array. This may be done in three different ways. For the closed system of Fig. 7, the three arrays are as follows:

~~CONFIDENTIAL~~

M \ K	1	2	3
M ₁	E ₁	E ₄	E ₂
M ₂	E ₃	E ₁	E ₄
M ₃	E ₄	E ₃	E ₁
M ₄	E ₂	E ₂	E ₃

M \ E	E ₁	E ₂	E ₃	E ₄
M ₁	1	3		2
M ₂	2		1	3
M ₃	3		2	1
M ₄		1,2	3	

E \ K	1	2	3
E ₁	M ₁	M ₂	M ₃
E ₂	M ₄	M ₄	M ₁
E ₃	M ₂	M ₃	M ₄
E ₄	M ₃	M ₁	M ₂

From the first of these message M₂ with key 3 yields cryptogram E₄. From the second M₁ is transformed into E₂ by key 3. No key transforms M₁ into E₃ and either 1 or 2 transforms M₄ into E₂. From the third E₃ is deciphered by key 2 to give M₃. All of these arrays and the line diagram contain equivalent information--from any one the others can be derived.

These arrays and diagrams only describe the set of transformation in the system. To specify the system the probabilities of various keys must also be given. This may be done by merely listing the keys with the associated probabilities. Similarly the message source is not completely specified until the probabilities of the various messages are given.

A more common way of describing a system is to describe the set of transformations by telling what operations one performs on the message for an arbitrary key to obtain the cryptogram. Similarly one defines implicitly the probabilities for various keys by describing how a key is chosen, or what we know of the enemy's habits of key choice. The probabilities for messages are implicitly determined by stating our a priori knowledge of the enemy's language habits, the tactical situation (which will influence the probable content of the message) and any special information we may have regarding the cryptogram.

10. Notation

The following notation will generally be followed.

- M = the message, also M₁, M_j, particular messages
- K = the key
- E = the enciphered message or cryptogram
- Ω_M = the set of all messages with associated probabilities, a probability space
- Ω_K = the set of keys with associated probabilities, also a probability space

Ω_E = the cryptogram space, also a probability space, since the probabilities in Ω_M and Ω_K induce probabilities in Ω_E , for each cryptogram.

m_i = the i^{th} letter of the message

e_i = the i^{th} letter of the cryptogram

k_i = the i^{th} letter of the key when it can be so described

Generally P stands for a probability. Conditional probabilities are indicated with subscripts. Thus

$P(M)$ = probability of message M

$P(E)$ = probability of cryptogram E

$P(K)$ = probability of key K

$P_M(E)$ = conditional probability of E if message M is chosen

$P_E(M)$ = conditional probability of M if cryptogram E is intercepted, i.e., the a posteriori probability of M if E is observed.

Q = equivocation, a concept to be defined precisely later which measures the uncertainty of some knowledge defined only by probabilities. We also have conditional equivocations, thus $Q_M(K)$ is the equivocation of key knowing the message.

$|K|$ = $-\sum P(K) \log P(K)$ the size of the key space

$|M|$ = $-\sum P(M) \log P(M)$ the size of the message space

$|E|$ = $-\sum P(E) \log P(E)$ the size of the cryptogram space

m = number of different keys

N = number of intercepted letters

R_0 = maximum information rate for a language

R = mean rate

$D = R_0 - R$ = redundancy of a language

T, R, S , etc. = secrecy systems

T_i, R_i, S_i , etc. = particular transformations of these systems

11. Some Examples of Secrecy Systems

In this section a number of examples of ciphers will be given. These will often be referred to in the remainder of the paper for illustrative purposes.

1. Simple Substitution Cipher.

In this cipher each letter of the message is replaced by a fixed substitute, usually also a letter. Thus the message

$$M = m_1 m_2 m_3 m_4 \dots$$

becomes

$$E = e_1 e_2 e_3 e_4 \\ = f(m_1) f(m_2) f(m_3) f(m_4) \dots$$

where the function $f(m)$ is function with an inverse. The key is a permutation of the alphabet (when the substitutes are letters) e.g. X G U A C D T B F H R S L M Q V Y Z W I E J O K. The first letter X is the substitute for A, G is the substitute for B, etc.

2. Transposition (Fixed Period d).

The message is divided into groups of length d and a permutation applied to the first group, the same permutation to the second group, etc. The permutation is the key and can be represented by a permutation of the first d integers. Thus for $d = 5$ we might have 2 3 1 5 4 as the permutation. This means that $m_1 m_2 m_3 m_4 m_5 m_6 m_7 m_8 m_9 m_{10} \dots$ becomes

$m_2 m_3 m_1 m_5 m_4 m_7 m_8 m_6 m_{10} m_9 \dots$ Sequential application

of two or more transpositions will be called compound transposition. If the periods are d_1, d_2, \dots, d_s it is clear that the result is a transposition of period d, where d is the least common multiple of $d_1, d_2, d_3, \dots, d_s$.

3. Vigenère, and Variations.

In this cipher the key consists of a series of d letters. There are written repeatedly below the message and two added modulo 26 (considering the alphabet numbered from A = 0 to Z = 25). Thus

$$e_1 = m_1 + k_1 \pmod{26}$$

where k_1 is of period d in the index 1.

For example with the key G A H we obtain

message	N O W I S T H E . . .
repeated key	G A H G A H G A . . .
cryptogram	T O D O S A N E . . .

The Vigenère of period 1 is called the Caesar cipher. It is a simple substitution in which each letter of M is advanced a fixed amount in the alphabet. This amount is the key, which may be any number from 0 to 25. The so-called Beaufort and

Variant Beaufort are similar to the Vigenère, and encipher by the equations

$$e_1 = k_1 - m_1 \pmod{26}$$

and

$$e_1 = m_1 - k_1 \pmod{26}$$

respectively. The Beaufort of period one is called the reversed Caesar cipher.

The application of two or more Vigenères in sequence will be called the compound Vigenère. It has the equation

$$e_1 = m_1 + k_1 + l_1 + \dots + s_1 \pmod{26}$$

where k_1, l_1, \dots, s_1 in general have different periods,

The period of their sum

$$k_1 + l_1 + \dots + s_1$$

as in compound transposition, is the least common multiple of the individual periods.

4. Vernam System.*

When the Vigenère is used with an unlimited key, never repeating, we have the Vernam system, with

$$e_1 = m_1 + k_1 \pmod{26}$$

the k_1 being chosen at random and independently among 0, 1, ..., 25. If the key is a meaningful text we have the "running key" cipher.

5. Bazeries Cylinder.

In this mechanical system 25 thick disks are used, each having a mixed alphabet stamped around the edge. These disks can be arranged in any order on a spindle, and the particular arrangement used constitutes the key. With the disks in their proper order, a message is enciphered by turning the disks so that the message appears on a line parallel to the axis of the spindle. Any other line of letters may then be chosen for the cryptogram. To decipher, the cryptogram is arranged on a line and the decipherer looks for another line which then makes sense.

*G. S. Vernam, "Cipher Printing Telegraph Systems for Secret Wire and Radio Telegraphic Communications," Journal Amer. Inst. of Elect. Eng., V, XLV, pp. 109-115, 1926.

CONFIDENTIAL

6. Digram, Trigram, and N-gram substitution.

Rather than substitute for letters one can substitute for digrams, trigrams, etc. General digram substitution requires a key consisting of a permutation of the 26^2 digrams. It can be represented by a table in which the row corresponds to the first letter of the digram and the column to the second letter, entries in the table being the substitutes (usually also digrams).

7. Interrupted Key Vigenère.

The Vigenère and its variations can be used with interrupted key. The sequence of key letters is started at irregularly spaced points. Thus, if the entire key sequence is X P G H F T R S, one can interrupt irregularly to get

X P G H F T X P G H F T R X P X P G . . .

The points of interruption can be determined in various ways (1). Whenever a certain letter occurs in the clear. (2). Whenever a certain letter occurs in the cryptogram. (3). An interrupting letter, say J, can be reserved as a signal at the encipherer interrupts the key at his discretion. (4). A signal is used and the decipherer locates the interruption by the appearance of meaningless text in the decipherment. In place of starting the key again at each interruption one can omit letters of it or reverse the direction of progress. There are many variations and combinations of these methods.

8. Single Mixed Alphabet Vigenère.

This is a simple substitution followed by a Vigenère.

$$e_1 = f(m_1) + k_1$$

$$m_1 = f^{-1}(e_1 - k_1)$$

The "inverse" of this system is a Vigenère followed by a simple substitution.

$$e_1 = g(m_1 + k_1)$$

$$m_1 = g^{-1}(e_1) - k_1$$

~~CONFIDENTIAL~~

9. Vigenère with Progressing Key.

The period of a Vigenère can be expanded by adding a fixed number t to the key at each appearance--thus the n th group is enciphered by the equation

$$e_i = m_i + k_i + nt$$

Also this can be varied by adding t and s alternately to the key, etc.

10. Matrix System.*

One method of n gram substitution is to operate on successive n -grams with a matrix having an inverse. The letters are assumed numbered from 0 to 25, making them elements of an algebraic ring. From the n -gram m_1, m_2, \dots, m_n of message, the matrix a_{ij} gives an n -gram of cryptogram

$$e_i = \sum_{j=1}^n a_{ij} m_j \quad i = 1, \dots, n$$

The matrix a_{ij} is the key, and deciphering is performed with the inverse matrix. The inverse matrix will exist if and only if the determinant $|a_{ij}|$ has an inverse element in the ring.

11. The Playfair Cipher.

This is a particular type of digram substitution governed by a mixed 25 letter alphabet written in a 5 x 5 square. (The letter J is often dropped in cryptographic work--it is very infrequent, and when it occurs can be replaced by I.) Suppose the key square is as shown below

L Z Q C P

A G N O U

R D M I F

K Y H V S

X B T E W

*See L. S. Hill, "Cryptography in an Algebraic Alphabet," American Math. Monthly, v. 36, No. 6, 1, 1929, pp.306-312, Also "Concerning Certain Linear Transformation Apparatus of Cryptography," v. 38, No. 3, 1931, pp.135-154.

The substitute for a digram AC, for example, is the pair c letters at the other corners of the rectangle defined by A and C, i.e. LO, the L taken first since it is above A. If digram letters are on a horizontal line as RI, one uses the letters to their right DF; RF becomes DR. If the letters on a vertical line, the letters below them are used. Thus becomes UW. If the letters are the same nulls may be used separate them or one may be omitted, etc.

12. Multiple Mixed Alphabet Substitution.

In this cipher there are a set of d simple substitutions which are used in sequence. If the period d is four

$m_1 m_2 m_3 m_4 m_5 m_6 \dots$

becomes

$f_1(m_1) f_2(m_2) f_3(m_3) f_4(m_4) f_1(m_5) f_2(m_6) \dots$

13. Autokey Cipher.

A Vigenère type system in which either the message itself or the resulting cryptogram is used for the "key" is called an autokey cipher. The encipherment is started with a "priming key" (which is the entire key in our sense) and continued with the message or cryptogram displaced by the length of the priming key as indicated below with the priming key COMET. The message used as "key",

MESSAGE SENDSUPPLIES ...

KEY COMETSENDSUP...

CRYPTOGRAM USZHLMTCOAYH

The cryptogram used as "key".

MESSAGE SENDSUPPLIES ...

KEY COMETUSZHL0H...

CRYPTOGRAM USZHL0H0STS ...

14. Fractional Ciphers.

In these, each letter is first enciphered into two or more letters or numbers and these symbols are somehow mixed (e.g. by transposition). The result may then be retranslated into the original alphabet. Thus using a mixed 25 letter alphabet for the key we may translate letters into two digit quinary numbers by the table

	0	1	2	3	4	
	0	L	Z	Q	C	P
1	A	G	N	O	U	
2	R	D	M	I	F	
3	K	Y	H	V	S	
4	X	B	T	E	W	

Thus B becomes 41. After the resulting series of numbers is transposed in some way they are taken in pairs and translated back into letters.

15. Codes.

In codes words (or sometimes syllables) are replaced by substitute letter groups. Sometimes a cipher of one kind or another is applied to the result.

12. Valuations of Secrecy Systems

There are a number of different criteria that should be applied in estimating the value of a proposed secrecy system. The more important of these are:

1. Amount of Secrecy.

There are some systems that are perfect--the enemy is no better off after intercepting any amount of material than before. Other systems, although giving him some information, do not yield a unique "solution" to intercepted cryptograms. Among the uniquely solvable systems, there are wide variations in the amount of labor required to effect this solution, and the amount of material that must be intercepted to make the solution unique.

2. Size of Key.

The key must be transmitted by non-interceptible means from transmitting to receiving ends. Sometimes it must be memorized. It is desirable then to have the key as small as possible.

3. Complexity of Enciphering and Deciphering Operations.

These should, of course, be as simple as possible. If they are done manually, complexity leads to loss of time, errors, etc. - If done mechanically, complexity leads to large expensive machines.

4. Propagation of Errors.

In certain types of secrecy systems an error of one letter in enciphering or transmission leads to a large amount of error in the deciphered text. The errors are spread out by the deciphering operation, causing the loss of much information and frequent need for repetition of the cryptogram. It is naturally desirable to minimize this error expansion.

5. Expansion of Message.

In some types of secrecy systems the size of the message is increased by the enciphering process. This undesirable effect may be seen in systems where one attempts to swamp out message statistics by the addition of many nulls, or where multiple substitutes are used. It also occurs in many "concealment" types of systems (which are not usually secrecy systems in the sense of our definition).

13. Equivalence Classes in the Key Space

It may happen that in a ciphering system two or more different keys, say keys 1, 2, and 7, are equivalent. By this we mean that for every M

$$T_1 M = T_2 M = T_7 M$$

These keys will not be considered as distinct but will be thrown into an equivalence class. It is clear that the cryptanalyst can never determine which particular one of these was used but only (at best) the class. The probability for the class is of course the sum of the probabilities of the different keys in the class.

As an example, in the Playfair cipher with the s given above, the following are equivalent key squares.

G H X P Y	E C I Z F
Z F E C I	N R D L O
L O N R D	V S Q T A
T A V S Q	W B M K U
K U W B M	X P Y G H

We can think of the possible equivalence classes in this case as arrangements of a 25 letter alphabet on a 5×5 square on an oriented torus. The number of different keys is not but $25!/5^2 = 24!$

When we say that two secrecy systems are the same mean that they consist of the same set of transformations with the same message and cryptogram space (range and domain) and the same probabilities for the different keys (after identical transformations are put in the same equivalence class).

14. The Algebra of Secrecy Systems

If we have two secrecy systems T and R we can combine them in various ways to form a new secrecy system. If T and R have the same domain (message space) we may form kind of "weighted sum,"

$$S = pT + qR$$

where $p + q = 1$. This operation consists of first making preliminary choice with probabilities p and q determining which of T and R is used. This choice is part of the key. After this is determined T or R is used as originally defined. The total key of S must specify which of T and R is used and which key of T (or R) is used.

If T consists of the transformations T_1, \dots, T_m with probabilities p_1, \dots, p_m and R consists of R_1, \dots, R_k with probabilities q_1, \dots, q_k then $S = pT + qR$ consists of the transformations $T_1, T_2, \dots, T_m, R_1, \dots, R_k$ with probabilities $pp_1, pp_2, \dots, pp_m, qq_1, qq_2, \dots, qq_k$ respectively.

More generally we can form the sum of a number of systems.

$$S = p_1 T + p_2 R + \dots + p_m U \quad \sum p_i = 1$$

We note that any system T can be written as a sum of fixed operations

$$T = p_1 T_1 + p_2 T_2 + \dots + p_m T_m$$

T_i being a definite enciphering operation of T corresponding to key choice i , which has probability p_i .

A second way of combining two secrecy systems is taking the "product", shown schematically in Fig. 8. Suppose T and R are two systems and the domain (language space) of T can be identified with the range (cryptogram space) of R. We can apply first R to our language and then T to the result of this enciphering process. This gives a resultant operation which we write as a product

$$S = TR$$

The key for S consists of both keys of T and R which are chosen according to their original probabilities and independently. Thus if the m keys of T are chosen with probabilities

$$p_1 p_2 \dots p_m$$

and the n keys of R have probabilities

$$p'_1 p'_2 \dots p'_n$$

then S has mn keys (at most; there may and often will be equivalence classes) with probabilities $p_i p'_j$. This type of product encipherment is often used; for example one follows a substitution by a transposition or a transposition by a Vigenère, or applies a code to the text and enciphers the result by substitution, transposition, fractionation, etc.

A more special type of product may be defined in case both T and R have keys of the same size which may be put in one-to-one correspondence with the same probabilities for corresponding keys. This may be called the "inner product", in contrast with the above which may be more completely described as an "outer product" (these names are derived from a rough analogy with the concepts of tensor analysis). In the inner product, written

$$S = T \circ R$$

and indicated schematically in Fig. 9, the same key (or corresponding keys) are used for both T and R chosen with the same probability.

For example one may construct a transposition cipher whose key is a permutation of the alphabet, each permutation being equally likely, and apply first this and then a substitution based on the same permutation. One also sees this situation in certain geometrical types of transposition ciphers where the text is written into a square and a permutation based on a key word applied first to the columns and then to the rows of the square.

It may be noted that multiplication (either kind) is not in general commutative, (we do not always have $RS = SR$) although in special cases such as substitution and transposition it is. Since it represents an operation it is definitionally associative. That is $R(ST) = (RS)T = RST$. Furthermore we have the laws

$$p(p'T + q'R) + qS = pp'T + pq'R + qS$$

(weighted associative law for addition)

$$T(pR + qS) = pTR + qTS$$

$$(pR + qS)T = pRT + qST$$

(right and left hand distributive laws)

and

$$p_1T + p_2T + p_3R = (p_1 + p_2)T + p_3R$$

Finally with regard to this algebraic structure of secrecy operations, we note that every closed secrecy system has an "inverse" T' obtained by interchanging the E and M spaces, with key probabilities the same, and

$$(TRS)' = S'R'T'$$

$$(pT + qR)' = pT' + qR'$$

Note that TT' is not in general the identity (this is the reason we do not write T^{-1}).

A system whose M and E spaces can be identified, a very common case as when letter sequences are transformed into letter sequences, may be termed endomorphic. An endomorphic system T may be raised to a power T^n .

A secrecy system T whose outer product with itself is equal to T , i.e. for which

$$T T = T$$

will be called idempotent. For example simple substitution transposition of period p , Vigenère of period p (all with a key equally likely) are idempotent.

The set of all endomorphic secrecy systems defined on a fixed message space constitute an "algebraic variety," that is, a kind of algebra, using the operations of addition and multiplication. In fact, the properties of addition and multiplication which we have discussed lead to the following result:

Theorem 1: The set of endomorphic ciphers with the same message space and the two combining operations of weighted addition and outer multiplication from a linear associative algebra with a unit element, apart from the fact that the coefficients in a weighted addition must be non-negative and sum to unity.

It should be emphasized that these combining operations of addition and multiplication apply to secrecy systems as a whole. The product of two systems TR should not be confused with the product of the transformations in the system T, R , which also appears often in this work. The former T is a secrecy system, i.e. a set of transformations with associated probabilities; the latter is a particular transformation. Further the sum of two systems $pR + qT$ is a system--the sum of two transformations is not defined. The systems T and R may commute without the individual T and R commuting, e.g. if R is a Beaufort system of a given period all keys equally likely,

$$R_1 R_2 + R_2 R_1$$

in general, but of course RR does not depend on its order; actually

$$RR = V$$

the Vigenère of the same period with random key. On the other hand, if the individual T and R of two systems T and R commute, then the systems commute.

It is rather surprising to find an algebraic variety with as much structure as a linear associative algebra in which

the elements have the complexity of ciphers. In Hilbert space theory, for example, one has a linear associative algebra, but the elements of the algebra are transformations. Here the elements are sets of transformations with a probability space associated with the transformation parameter.

These combining operations give us ways of constructing many new types of secrecy systems from certain ones, such as the examples given. We may also use them to describe the situation facing a cryptanalyst when attempting to solve a cryptogram of unknown type. He is, in fact, solving a secrecy system of the type

$$T = p_1 A + p_2 B + \dots + p_r S + p' X \sum p = 1$$

where the A, B, ..., S are known types of ciphers, with the p_i their a priori probabilities in this situation, and $p' X$ corresponds to the possibility of a completely new unknown type of cipher.

In weighted addition the key size of the result is given by

$$\begin{aligned} |K| &= - \sum_i p p_i \log p p_i - \sum q p_i'' \log q p_i'' \\ &= p |K_1| + q |K_2| - (p \log p + q \log q) \\ &= p |K_1| + q |K_2| + |K_3| \end{aligned}$$

i.e. the weighted mean of the two keys plus the size of the p, q key. This is only in case there are no equivalences; if there are it will always be less.

For the outer product the key size is

$$|K| \leq |K_1| + |K_2|$$

with equality only when there are no equivalences. In the inner product

$$|K| \leq |K_1| - |K_2|$$

with equality under the same condition.

15. Pure and Mixed Ciphers

Certain types of ciphers, such as the simple substitution, the transposition of a given period, the Vigené of a given period, the mixed alphabet Vigenère, etc. (all with each key equally likely) have a certain homogeneity w respect to key. Whatever the key, the enciphering, deciphering and decrypting processes are essentially the same. This may be contrasted with the cipher

$$p S + q T$$

where S is a simple substitution and T a transposition of given period. In this case the entire system changes for enciphering, deciphering and decryptment, depending on whe the substitution or transposition was used.

The cause of the homogeneity in certain ciphers stems from the group property--we notice that in the above amples of homogeneous ciphers the product of any two transformations in the set T_1, T_j is equal to a third transforme T_k in the set, while $T_1 S_j$ does not equal any transformat in the cipher

$$p S + q T$$

which contains only substitutions and transpositions, no products.

We might define a "pure" cipher, then, as one wh T_1 formed a group. This, however, would be too restrictiv since it requires that the E space be the same as the M sp i.e. that the system be endomorphic. The frectional trans position is as homogeneous as the ordinary transposition w out being endomorphic. The proper definition is the follo A cipher T is pure if for every T_1, T_j, T_k there is a T_s s that

$$T_1 T_j^{-1} T_k = T_s$$

and every key is equally likely. Otherwise the cipher is The systems of Fig. 7 are mixed. Fig. 10 is pure if all k are equally likely.

Theorem 2: In a pure cipher the operations $T_1^{-1} T_j$ which transform the message space into itself form group whose order is m, the number of differen keys.

For

$$T_j^{-1} T_k T_k^{-1} T_j = I$$

so that each element has an inverse, also the associative law is true since these are operations, and the group property follows from

$$T_i^{-1} T_j T_k^{-1} T_l = T_s^{-1} T_k T_k^{-1} T_l = T_s^{-1} T_l$$

using our assumption that $T_i^{-1} T_j = T_s^{-1} T_k$ for some s .

The operation $T_i^{-1} T_j$ means, of course, encipher the message with key j and then deciphering with key i which brings us back to the message space. If T is endomorphic i.e. the T_i themselves transform the space Ω_M into itself is the case with most ciphers, where both the message space and the cryptogram space consist of sequences of letters and the T_i are a group and equally likely, then T is pure since

$$T_i T_j^{-1} T_k = T_i T_r = T_s$$

Theorem 3: The outer product of two pure ciphers which commute is pure.

For if T and R commute $T_i R_j = R_l T_m$ for every i, j with suitable l, m , and

$$\begin{aligned} T_i R_j (T_k R_l)^{-1} T_m R_n &= T_i R_j R_l^{-1} T_k^{-1} \\ &= R_u R_v^{-1} R_w T_r T_s \\ &= R_h T_g \end{aligned}$$

The commutation condition is not necessary, however, for product to be a pure cipher.

A system with only one key, i.e. a single defining operation T_1 , is pure since the only choice of indices is

$$T_1 T_1^{-1} T_1 = T_1.$$

Thus the expansion of a general cipher into a sum of such simple transformations also exhibits it as a sum of pure ciphers.

An examination of the example of a pure cipher shown in Fig. 5 discloses certain properties. The messages fall into certain subsets which we will call residue classes and the possible cryptograms are divided into corresponding residue classes. There is at least one line from each message in a class to each cryptogram in the corresponding class and no line between classes which do not correspond. The number of messages in a class is a divisor of the total number of keys. The number of lines "in parallel" from a message M to a cryptogram in the corresponding class is equal to the number of keys divided by the number of messages in the class containing the message (or cryptogram). It is shown in the appendix that these hold in general for pure cipher. Summarized in a more formal statement we have

Theorem 4: In a pure system the messages can be divided into a set of "residue classes" C_1, C_2, \dots, C_s and the cryptograms into a corresponding set of residue classes C'_1, C'_2, \dots, C'_s with the following properties

- (1) The message residue classes are mutually exclusive and collectively contain all possible messages. Similarly for the cryptogram residue classes.
- (2) Enciphering any message in C_i with any key produces a cryptogram in C'_i .¹ Deciphering any cryptogram in C'_i with any key leads to a message in C_i .
- (3) The number of messages in C_i , say ϕ_i , is equal to the number of cryptograms in C'_i and is a divisor of k the number of keys.
- (4) Each message in C_i can be enciphered into each cryptogram in C'_i by exactly $\frac{k}{\phi_i}$ different keys. Conversely ϕ_i for decipherment.

The importance of the concept of a pure cipher (the reason for the name) lies in the fact that for them all keys are essentially the same. Whatever key is used for a particular message, the a posteriori probabilities of a messages are identical. To see this, note that two different keys applied to the same message lead to two cryptograms in the same residue class, say C_1 . The two cryptograms therefore could each be deciphered¹ by $\frac{k}{\phi_1}$ keys into each message

in C_1 , and into no other possible messages. All keys being equally likely the a posteriori probabilities of various messages are thus

$$P_E(M) = \frac{P(M) P_M(E)}{\sum P(M) P_M(E)} = \frac{P(M)}{P(C_1)}$$

where M is in C_1 , E is in C_1 and the sum is over all messages in C_1 . If E and M are not in corresponding residue classes $P_E(M) = 0$. Similarly it can be shown that the a posteriori probabilities of the different keys are the same in value. These values are associated with different keys when a different key is used. The same set of values of $P_E(K)$ have undergone a permutation among the keys. Thus we have the result

Theorem 5: In a pure system the a posteriori probabilities of various messages $P_E(M)$ are independent of the key that is chosen. The a posteriori probabilities of the keys $P_E(K)$ are the same in value but undergo a permutation with a different key choice.

Roughly we may say that any key choice leads to the same cryptanalytic problem in a pure cipher. Since the different keys all result in cryptograms in the same residue class this means that all cryptograms in the same residue class are cryptanalytically equivalent--they lead to the same a posteriori probabilities of messages and, apart from a permutation, the same probabilities of keys.

As an example of this, simple substitution with all keys equally likely is a pure cipher. The residue class corresponding to a given cryptogram E is the set of all cryptograms that may be obtained from E by operations T_j . In this case T_j is itself a substitution and hence an operation on E gives another member of the same residue class. Thus if the cryptogram is

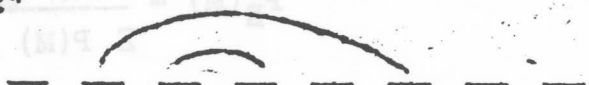
$E = X C P P G C F Q$

then

$E_1 = R D H H G D S N$

$E_2 = A B C C D B E F$

etc. are in the same residue class. It is obvious in this case that these cryptograms are essentially equivalent. All that is of importance in a simple substitution with random key is the pattern of letter repetitions, the actual letters being dummy variables. Indeed we might dispense with them entirely indicating the pattern of repetitions in E as follows:*



This notation describes the residue class but eliminates a information as to the specific member of the class. Thus leaves precisely that information which is cryptanalytical pertinent. This is related to one method of attacking simple substitution ciphers--the method of pattern words.

In the Caesar type cipher only the first differences mod 26 of the cryptogram are significant. Two cryptograms with the same Δe_i are in the same residue class. One breaks this cipher by the simple process of writing down the 26 members of the message residue class and picking out the one which makes sense.

The Vigenère of period d with random key is an example of a pure cipher. Here the message residue class consists of all sequences with the same first differences letters separated by distance d as the cryptogram. For $d = 3$ the residue class is defined by

$$m_1 - m_4 = e_1 - e_4$$

$$m_2 - m_5 = e_2 - e_5$$

$$m_3 - m_6 = e_3 - e_6$$

$$m_4 - m_7 = e_4 - e_7$$

⋮

*Suggested by a notation used by Quine in Symbolic Logic.

where $E = e_1, e_2, \dots$ is the cryptogram and m_1, m_2, \dots is any M in the corresponding residue class.

In the transposition cipher of period d with random key, the residue class consists of all arrangements of the e_i in which no e_i is moved out of its block of length d , and any two e_i at a distance d remain at this distance. This is used in breaking these ciphers as follows. The cryptogram is written in successive blocks of length d , one under another as below ($d = 5$):

e_1	e_2	e_3	e_4	e_5
e_6	e_7	e_8	e_9	e_{10}
e_{11}	e_{12}	.	.	.
.

The columns are then cut apart and rearranged to make sense. When the columns are cut apart, the only information remaining is the residue class of the cryptogram.

Theorem 6: If T is pure then $T_i T_j^{-1} T = T$ where T_i, T_j are any two transformations of T . Conversely if this is true for any T_i, T_j in a system T then T is pure.

The first part of this theorem is obvious from the definition of a pure system. To prove the second part we note first that if $T_i T_j^{-1} T = T$ then $T_i T_j^{-1} T_s$ is a transformation of T . It remains to show that all keys are equiprobable.

We have $T = \sum_s p_s T_s$ and

$$\sum_s p_s T_i T_j^{-1} T_s = \sum_s p_s T_s$$

the term in the left hand sum with $s = j$ yields $p_j T_i$. The only term in T_i on the right is $p_i T_i$. Since all coefficients are non negative it follows that

$$p_j \leq p_i$$

The same argument holds with i and j interchanged and consequently

$$p_j = p_i$$

and T is pure. Thus the condition that $T_i T_j^{-1} T = T$ might be used as an alternative definition of a pure system.

The property of purity in a system is connected with idempotence. Thus consider the system $S = T T'$ where T is pure. We have

$$T_i T_j^{-1} T_s T_r^{-1} = T_i T_l^{-1} T_r T_r^{-1} = T_i T_l^{-1}$$

so that the transformations of S^2 are the same as those of S , and since both S and S^2 are pure we have

$$S = S^2$$

Theorem 7: If T is pure $S = T T'$ is pure and $S^2 = S$.

An endomorphic system T which satisfies the condition $T_i T_j = T_s$ (but not necessarily with all key probabilities equal) can be shown to approach a pure cipher on raising to a high power, namely the one with the same transformations, but with all probabilities equalized. In fact the probabilities for T^{n+1} are derived from those for T^n by a Markoff process, of a special type due to the group property. This special type always approaches the limit of equalized probabilities. This same argument applies more generally. We have

Theorem 8: Let T be any endomorphic cipher. If T^n approaches any limit at all, which will necessarily occur if all the transformations of T^n lie in a finite set (no matter how large n) and the transformations of T include the identity then this limit will be a pure cipher.

As an example consider the cipher

$$R = p T + q S$$

where T is transposition with random key and S substitution with random key. We have

$$S^2 = S$$

$$T^2 = T$$

$$S T = T S$$

and hence any product of T 's and S 's such as $T S T T S S$ reduces to $S T$. Thus

$$R^n = p^n T + q^n S + (1 - p^n - q^n) S T$$

As $n \rightarrow \infty$ the first two terms approach zero and

$$\lim_{n \rightarrow \infty} R^n = S T$$

The concepts of pure and mixed languages and pure and mixed ciphers have an application in practical cryptanalysis, if we interpret them somewhat loosely. When a cryptographer starts work on a cryptogram, his first job is to determine the original language. Approximately then he is determining the pure component of the general language space

$$L = p_1 L_1 + p_2 L_2 + \dots + p_n L_n$$

where L_1 say is English, L_2 German, etc. Of course these are not pure but the different components of them are fairly close together in statistical structure.

The second thing a cryptographer does is to determine the "type" of cipher that was used--usually this is about the same as finding the pure component in the general cipher system

$$R = p_1 S + p_2 T + p_3 V + \dots$$

where S say is simple substitution, T is transposition, etc. A Vigenère V of unknown period is not a pure cipher but the decomposition

$$V = p_1 V_1 + p_2 V_2 + p_3 V_3 + \dots$$

where V_i is of period i , is into pure components (if all keys are equally likely for any period). In solving a Vigenère the first problem is to determine the period. The same is true in transposition.

The reason for this initial isolation of pure or nearly pure language and cipher is that only then can a simple meaningful statistical analysis be carried out.

16. Involutory Systems

If every transformation in a system T is its own inverse, i.e. if

$$T_i T_i = I$$

for every i , the system will be called involutory. Such systems are important practically since the enciphering and deciphering operations are then identical. This leads to simplified instructions to cryptographic clerks in manual operation, or in mechanical cases the same machine with the same key setting may be used for both operations.

Examples: In simple substitution we may limit our transformations to those in which when letter θ is the substitute for ϕ , ϕ is the substitute for θ . Another example is the Beaufort cipher.

If T is involutory, so is the system whose operations are

$$S_j T_i S_j^{-1}$$

since

$$S_j T_i S_j^{-1} (S_j T_i S_j^{-1}) = S_j T_i S_j^{-1} S_j T_i S_j^{-1} = I$$

17. Similar and Weakly Similar Systems

Two secrecy systems R and S will be said to be similar if there exists a transformation A having an inverse A^{-1} such that

$$R = A S$$

This means that enciphering with R is the same as enciphering with S and then operating on the result with the transformation A . If we write $R \approx S$ to mean R is similar to S then it is clear that $R \approx S$ implies $S \approx R$. Also $R \approx S$ and $S \approx T$ imply $R \approx T$ and finally $R \approx R$. These are summarized in mathematical terminology by saying that similarity is an equivalence relation.

The cryptographic significance of similarity is that if $R \approx S$ then R and S are equivalent from the cryptanalytic point of view. Indeed if a cryptanalyst intercepts a cryptogram in system S he can transform it to one in system R by merely applying the transformation A to it. A cryptogram in system R is transformed to one in S by applying A^{-1} . If R and S are applied to the same language or message space, there is a one-to-one correspondence between the resulting cryptograms. Corresponding cryptograms give the same distribution of a posteriori probabilities for all messages.

If one has a method of breaking the system R then any system S similar to R can be broken by reducing to R through application of the operation A. This is a device that is frequently used in practical cryptanalysis.

Examples: As a trivial example, simple substitution where the substitutes are not letters but arbitrary symbols is similar to simple substitution using letter substitutes. A second example is the Caesar and the reversed Caesar type ciphers. The latter is sometimes broken by first transforming into a Caesar type. The Vigenère, Beaufort and Variant Beaufort are all similar, when the key is random. The "autokey" cipher primed with the key $K_1 K_2 \dots K_n$ is similar to a Vigenère type with the key alternately added and subtracted mod 26. The transformation A in this case is that of "deciphering" the autokey with a series of d A's for the priming key.

Two systems R and S are weakly similar if there exist two transformations A and B having inverse A^{-1} and B^{-1} with

$$R = A S B$$

This means that system R is the same as applying first B to the language, then S, and finally A. This relation is also an equivalence relation.

Finding a method of solution for system R with language L is equivalent to finding a solution for S with language B L.

We may note that if R is pure and S is weakly similar to R then S is pure. This follows from

$$R_i R_j^{-1} R_k = R_l$$

$$R_i = A S_i B$$

$$R_j^{-1} = B^{-1} S_j^{-1} A^{-1}$$

$$R_k = A S_k B$$

where we assume corresponding transformations in R and S to have the same subscripts. Hence

~~CONFIDENTIAL~~

$$R_i R_j^{-1} R_k = A S_i S_j^{-1} S_k B = R_l$$

$$S_i S_j^{-1} S_k = A^{-1} R_l B^{-1}$$

$$= S_l$$

and S is therefore pure.

$$R = A S B$$

language B is equivalent to finding a solution for S with language A. This means that system R is the same as applying first B to the language, then S, and finally A. This relation is also an equivalence relation.

We may note that if R is pure and S is weakly similar to R then S is pure. This follows from

$$R_1 R_2^{-1} R_3 = R_4$$

$$R_1 = A S_1 B$$

$$R_2 = A S_2 B$$

$$R_3 = A S_3 B$$

where we assume corresponding transformations for R and S to have the same alphabets. Hence

PART II

Theoretical Secrecy

Introduction

We now consider problems connected with the "theoretical secrecy" of a system. How immune is a system to cryptanalysis when the cryptanalyst has unlimited time and manpower available for the analysis of cryptograms? Does a cryptogram have a unique solution (even though it may require an impractical amount of work to find it) and if not how many reasonable solutions does it have? How much text in a given system must be intercepted before the solution becomes unique? Are there systems which never become unique in solution no matter how much enciphered text is intercepted? Are there systems for which no information whatever is given to the enemy no matter how much text is intercepted?

18 Perfect Secrecy

Let us suppose the possible messages are finite in number $M_1 \dots M_N$ and have a priori probabilities $P(M_1), \dots, P(M_N)$, and that these are enciphered into the possible cryptograms $E_1 \dots E_M$ by

$$E = T_1 M.$$

The cryptanalyst intercepts a particular E and can then calculate the a posteriori probabilities for the various messages, $P_E(M)$. It is natural to define perfect secrecy by the condition that for all E , the a posteriori probabilities are equal to the a priori probabilities independently of the values of these. In this case, intercepting the message has given the cryptanalyst no information.* Any action of his which depends on the information contained in the cryptogram cannot be altered, for all of his probabilities as to what the cryptogram contains remain unchanged. On the other hand, if the condition is not satisfied there will exist situations in which the enemy has certain a priori probabilities, and certain key and messages are chosen where the enemy's probabilities do change. This in turn may affect his actions and thus perfect secrecy has not been

*A purist might object that the enemy has obtained a bit of information in that he knows a message was sent. This may be answered by having among the messages a "blank" corresponding to "no message." If no message is originated the blank is enciphered and sent as a cryptogram. Then even this modicum of remaining information is eliminated,

obtained. Hence the definition given is necessarily required by our ideas of what perfect secrecy should mean.

A necessary and sufficient condition for perfect secrecy can be found as follows. We have by Bayes' theorem

$$P_E(M) = \frac{P(M) P_M(E)}{P(E)}$$

and this must equal $P(M)$ for perfect secrecy. Hence either $P(M) = 0$, a solution that must be excluded since we demand the equality independent of the values of $P(M)$, or

$$P_M(E) = P(E)$$

for every M and E . Conversely if $P_M(E) = P(E)$ then

$$P_E(M) = P(M)$$

and we have perfect secrecy. Thus we have the result:

Theorem 9: A necessary and sufficient condition for perfect secrecy is that

$$P_M(E) = P(E)$$

for all M and E . That is $P_M(E)$ must be independent of M .

The probability of all keys that transform M_i into a given cryptogram E is equal to that of all keys transforming M_j into the same E .

Now there must be as many E 's as there are M 's, since fixing i , T_i gives a one-to-one correspondence between all the M 's and some of the E 's. For perfect secrecy $P_M(E) = P(E) \neq 0$ for any of these E 's and any M . Hence there is at least one key transforming any M into any of these E 's. But all the keys from a fixed M to different E 's must be different, and therefore the number of different keys is at least as great as the number of M 's. It is possible to obtain perfect secrecy with no more, as one shows by the following example. Let the M_i be numbered 1 to n and the E_i the same, and using n keys let

$$T_i M_j = E_s$$

where $s = i + j \pmod{n}$. In this case we see that $P_E(M) = \frac{1}{n} = P(E)$ and we have perfect secrecy. An example is shown in Fig. 11 with $n = 5$.

These perfect systems in which the number of cryptograms, the number of messages, and the number of keys are all equal are characterized by the properties that (1) each M is connected to each E by exactly one line, (2) all keys are equally likely. Thus the three matrix representations of the system "latin squares".

We have then concealed completely an amount of information at most $\log n$ with a size of key $\log n$. This is the first example of a general principle which we will often see, that there is a limit to what can obtain with a given key size--the amount of uncertainty we can introduce into the solution of a cryptogram cannot be greater than the key size. Here we have concealed all the information but the key size is as large as a message space.

We now consider the case where $|M|$ is infinite; in suppose the message generated as an unending sequence of letters by a Markoff process. The maximum rate of this source is R_0 . It is clear from our results above that no finite key will give perfect secrecy. We suppose then that the key source generates key also in the same manner, i.e. as an infinite sequence of letters with a mean rate R_K . Suppose that only a certain length key L_K is needed to encipher and decipher a length L_M of message.

Theorem 10: For perfect secrecy (when the a priori probabilities of various messages can be anything), for large L

$$R_0 L_M \leq R_K L_K$$

and the rate $(R_K + \epsilon)$ is asymptotically sufficient.

This may be proved by the same method (essentially the finite case. This case is realized by the Vernam system).

These results have been deduced on the basis of unknown or arbitrary a priori probabilities for the messages. The key required for perfect secrecy depends then on the total number of possible messages, or on the maximum rate R_0 of the message source.

One would suspect that if the message space has finite known statistics, so that it has a definite mean rate R of generating information, then the amount of key needed could be reduced in an average sense in just this ratio $\frac{R}{R_0}$, and this

indeed true. In fact the message can be passed through a transformer which transforms it into a normal form and reduces the

expected length in just this ratio, and then a Vernam system may be applied to the result. Evidently the amount of key per letter of message is statistically reduced by a factor

$\frac{R}{R_0}$ and in this case the key source and information source

just matched--an alternative of key conceals an alternative information. It is easily seen also, by the methods used in "Information" paper that this is the best that can be done.

Theorem 11: Perfect secrecy (omitting the condition of independence of a priori probabilities) for a source with fixed statistics and a rate R of generating information can be achieved with a key source which generates at the rate $(R + \epsilon) \frac{L_M}{L_K}$ where L_M and L_K are message

and key lengths which correspond. A rate less than $R \frac{L_M}{L_K}$ is insufficient.

Perfect secrecy systems have a place in the practical picture--they may be used either where the greatest importance is attached to complete secrecy--e.g. correspondence between the highest levels of command, or in cases where the number of possible messages is small. Thus, to take an extreme example if only two messages "yes" or "no" were anticipated a perfect system would be in order, with perhaps the transformation

M	K	
	A	B
yes	0	1
no	1	0

The disadvantage of perfect systems for large correspondence systems is, of course, the equivalent amount of key that must be sent. In succeeding sections we consider what can be achieved with smaller key size, in particular with finite keys.

19. Equivocation

Let us suppose that a simple substitution cipher has been used on English text and that we intercept a certain N letters, of the enciphered text. For N fairly large, more than say 50 letters, there is nearly always a unique solution to the cipher; i.e. a single good English sequence which trans

into the intercepted material by a simple substitution. With smaller N , however, the chance of more than one solution is greater; with $N = 15$ there will generally be quite a number of possible fragments of text that would fit, while with $N = 8$ a good fraction (of the order of $1/8$) of all reasonable English sequences of that length are possible, since there is seldom more than one repeated letter in the 8. With $N = 1$ any letter is clearly possible and has the same a posteriori probability as its a priori probability. For one letter the system is perfect.

This happens generally with solvable ciphers. When any material is intercepted we can imagine the a priori probabilities attached to the various possible messages, and also to the various keys. As material is intercepted, the cryptanalyst calculates the a posteriori probabilities, and as N increases the probabilities of certain messages increase, and of most decrease, until finally only one is left, which has a probability nearly one, while the total probability of all others is nearly zero.

This calculation can actually be carried out for simple systems. Table 1 shows the a posteriori probabilities for a Caesar type cipher applied to English text, with the key chosen at random from the 26 possibilities. To enable the use of standard letter digram and trigram frequency tables the message has been started at a random point (by opening a book and putting a pencil down at random on the page). The message selected this way begins "creases to . . ." starting inside the word "creases". If the message were to start with the beginning of a sentence a different set of probabilities must be used, corresponding to the frequencies of letters, digram, etc., at the beginning of sentences.

The Caesar with random key is a pure cipher and the particular key chosen does not affect the a posteriori probabilities. To determine these we need merely list the possible decipherments by all keys and calculate their a priori probabilities. The a posteriori probabilities are these divided by their sum. These possible decipherments are found by the standard process of "running down the alphabet" from the message and are listed at the left. These form the residue class of the message. For one intercepted letter the a posteriori probabilities are equal to the a priori probabilities for letters, as shown in the column headed $N = 1$. For two intercepted letters the probabilities are those for digram adjusted to sum to unity and these are shown in the column $N = 2$.

Table 1

A Posteriori Probabilities for a Caesar Type Cryptogr

Decipherments	N = 1	N = 2	N = 3	N = 4	N =
C R E A S	.032	.015	.111	.55	1
D S F B T	.036	.068			
E T G C U	.123	.170			
F U H D V	.023	.023			
G V I E W	.016				
H W J F X	.051	.015			
I X K G Y	.072				
J Y L H Z	.001				
K Z M I A	.005				
L A N J B	.040	.072	.250	.01	
M B O K C	.020	.019	.022	.01	
N C P L D	.072	.066			
O D Q M E	.079	.034			
P E R N F	.023	.085	.438	.43	
Q F S O G	.002				
R G T P H	.060	.013			
S H U Q I	.066	.064	.005		
T I V R J	.096	.272	.166		
U J W S K	.030				
V K X T L	.009				
W L Y U M	.020	.008	.005		
X M Z V N	.002				
Y N A W O	.019	.006			
Z O B X P	.001				
A P C Y Q	.080	.066			
B Q D Z R	.016				
Q (digits) =	1.248	.999	.602	.340	0

Trigram frequencies have also been tabulated and these are in column N = 3. For four and five letter sequences probabilities were obtained by multiplication from trigram frequencies since approximately

$$p(ijkl) = p(ijk) p_{jk}(l)$$

Note that at three letters the field has narrowed to four messages of fairly high probability, the others being small in comparison. At four there are two possibilities five just one, the correct decipherment.

In principle this could be carried out with any but unless the key is very small the number of possibilities so large that the work involved prohibits the actual calculation.

This set of a posteriori probabilities describes the cryptanalyst's knowledge of the message and key gradually becomes more precise as enciphered material is obtained. description, however, is much too involved and difficult to obtain for our purposes. What is desired is a simplified description of this approach to uniqueness of the possible solutions.

We will first define a quantity Q called the "equivocation" which measures in an average way the uncertainty of the solution, or how far it is from uniqueness. Suppose that a certain cryptogram E of N letters has been intercepted. A cryptanalyst can in principle calculate the a posteriori probabilities by the use of Bayes' theorem. Thus

$$P_E(M) = P(M) P_M(E)/P(E).$$

Similarly the probabilities for various keys, after E has been intercepted are given by

$$P_E(K) = P(K) P_K(E)/P(E).$$

The equivocation of the message should measure how far spread out these probabilities $P_E(M)$ are; how far they are from being concentrated at one message. In line with general principles of measuring such dispersion, as in the case of choice, uncertainty, and generating information, we define the equivocation of the message when E has been intercepted

$$Q(M) = - \sum_M P_E(M) \log P_E(M)$$

the summation being over all possible messages. Similarly the equivocation in key when E is intercepted is given by

$$Q(K) = - \sum_K P_E(K) \log P_E(K)$$

The same general arguments used to justify our measure of information rate may be used here, to justify the equivocation measure. We note that equivocation zero requires that one message (or key) have probability one, all others zero. Equivocation is measured in the same units as information, i.e. alternative digits, etc., according as the logarithmic base is 2, 10, etc. In fact, equivocation is almost identical with information, the difference being one of point of view. In information we use the notion of how much freedom we have in choosing one element from a set with certain probabilities--in equivocation we use the size of the uncertainty of our knowledge of what was chosen when the probabilities have certain values.

Although any one number can hardly be expected to describe the set $P_E(M)$ perfectly for all purposes, I think the definition here does as well as any single statistic can. Some of the theorems which follow indicate the mathematical "nature" of this particular measure.

The values of equivocation for the Caesar type cipher program considered above have been calculated and are given in the last row of Table 1. This is the Q for both key and message, the two being equal in this case.

The definitions given above involve a particular intercepted message E , and are the equivocations for that intercepted message. We wish, however, to find a measure of the equivocation for the system as a whole, which will describe this progression toward uniqueness as N increases in an average sort of way. To do this we form a weighted average of the equivocations for each particular intercepted message E , weighting in accordance with the probabilities of getting the E in question. This will be called the mean equivocation of the system, or where there is no chance of confusion with the narrower equivocation for a particular E , we abbreviate to merely the equivocation. The mean equivocation of message is

$$Q(M) = - \sum_{M,E} P(E) P_E(M) \log P_E(M)$$

the summation being over all M and all E . Since

$$P(E) P_E(M) = P(E, M)$$

the probability of getting both E and M , we can write this

$$Q(M) = - \sum P(M,E) \log P_E(M) = - \sum P(M,E) \log P(M) \frac{P_M(E)}{P(E)}.$$

Similarly

$$Q(K) = - \sum P(K,E) \log P(K) \frac{P_K(E)}{P(E)} .$$

Either of these mean equivocations is a theoretical measure of the secrecy value of the system. We say theoretical since even when the equivocation is zero, which corresponds to no uncertainty as to the message, it may require a tremendous amount of labor to locate the particular message where the probability is one. It might, for example, be necessary to try every possible K in succession until one was found that transformed the intercepted E into reasonable text in the language. Thus the system would be practically very good, but theoretically so. The equivocation may be said to measure the degree of secrecy when the cryptanalyst has unlimited time and energy.

The equivocation is, of course, a function of N, the number of letters intercepted. The functions $Q(K,N)$ and $Q(M,N)$ will be called the equivocation characteristics of the system.

The following data will be helpful in forming a picture of what small values of equivocation represent.

An equivocation of .1 alternative would result if (1) 9 times in 10 there was no uncertainty as to M, the tenth time two M's were equally probable, or (2) if every time there were two possibilities one with probability .983, the other with probability .017, or (3) if 99 times in 100 there was no uncertainty, the 100th time 1000 equally likely possibilities.

An equivocation of .01 would result (1) if every time there were two possibilities one with probability .999, the other with probability .001, or (2) if 99 times in 100 there is no uncertainty, the other time two equally likely possibilities, or (3) if 999 times in 1000 there is no uncertainty, the other time 6 or 7 equally likely possibilities.

20. Properties of Equivocation

Equivocation may be shown to have a number of interesting properties, most of which fit into our intuitive picture of how such a quantity should behave. We may first show, by example, the somewhat surprising fact, that after a cryptanalyst has intercepted certain special E's, his equivocation as to the message may be greater than before he intercepted anything. The intercepted material has increased his ignorance of what happened! Suppose there are only two messages M_1 and M_2 with a priori probabilities p and q, and that a simple substitution

is used according to the following table, the two keys K_1 and K_2 also having the a priori probabilities p and q .

	K_1	K_2
M_1	E_2	E_1
M_2	E_1	E_2

Before the interception, the equivocation of both key and message is $-(p \log p + q \log q)$, which is less than one alternative if $p \neq q$. If $p \gg q$ there is little uncertainty as to which message and key will be chosen, M_1 and K_1 . Now suppose he intercepts E_1 .

The a posteriori probabilities of both keys and both messages are easily seen to be $1/2$, and hence the equivocation for both key and message is one alternative, greater than before. On the other hand, if E_2 is intercepted, the more probable event, the equivocation for both key and message decreases, more than enough to compensate for the other increase, and the mean equivocation of both key and message decreases. This is a general property of all secrecy systems.

Theorem 12: The mean equivocation of key, $Q_K(N)$ is a non-increasing function of N . The mean equivocation of the first A letters of the message is a non-increasing function of the number N which have been intercepted. If N letters have been intercepted, the equivocation of the first N letters of message is less than or equal to that of the key. These may be written

$$\begin{aligned} Q_K(S) &\leq Q_K(N) & S &\geq N \\ Q_M(M) &\leq Q_M(N) & M &\geq N \\ Q_M(N) &\leq Q_K(N) \end{aligned}$$

The qualification regarding A letters in the second result of the theorem is so that the equivocation will not be calculated with respect to the amount of message that has been intercepted. If it is, the message equivocation may (and usually does) increase for a time, due merely to the fact that more letters stand for a larger possible range of messages. The results of the theorem are what we might hope from a good measure of equivocation, since we would hardly expect to be worse off on the average after intercepting material than before. The fact that they can be proved gives additional justification to our definition.

The results of this theorem can be proved by a substitution in the property 6 of section 1. Thus to prove the first or second we have for any chance events A and B

$$Q(B) \geq Q_A(B)$$

If we identify B with the key (knowing the first S letters of cryptogram) and A with the remaining N - S letters we obtain the first result. Similarly identifying B with the message gives the second result. The last result follows from

$$Q(M) \leq Q(K) + Q_K(M)$$

and the fact that $Q_K(M) = 0$ since K uniquely determines M.

Theorem 13: $Q(K) = |M| - |E| + |K|$

$$Q(M) = |M| - |E| + |H|$$

where

$$|H| = - \sum_{M,E} P(M,E) \log P_M(E)$$

We have.

$$Q(K) = - \sum_{E,K} P(E) P_E(K) \log P_E(K)$$

$$P_E(K) = \frac{P(K) P_K(E)}{P(E)}$$

Hence

$$Q(K) = - \sum P(K) P_K(E) \log P(K) - \sum P(K) P_K(E) \log P_K(E) + \sum P(K) P_K(E) \log P(E)$$

Summing the first term on E gives $-\sum P(K) \log P(K) = |K|$. In the second term $P_K(E)$ is $P(M)$, the unique M that gives E with key K. Summing on K then gives $-\sum P(M) \log P(M) = |M|$. The third term is $\sum P(E) \log P(E) = |E|$.

The second equation in the theorem is proved by the same method.

$$\begin{aligned}
 Q(M) &= - \sum P(E) P_E(M) \log P_E(M) \\
 &= - \sum P(M) P_M(E) \log \frac{P(M) P_M(E)}{P(E)} \\
 &= - \sum P(M) P_M(E) \log P(M) - \sum P(M) P_M(E) \log P_M(E) \\
 &\quad + \sum P(M) P_M(E) \log P(E) \\
 &= |M| - |E| - \sum P(M) P_M(E) \log P_M(E)
 \end{aligned}$$

The last term here may be interpreted as follows. Group together all the different keys that transform a fixed M into the same E, giving the total probability to the group, which will be $P_M(E)$. The last term is the average size of this group space weighted according to the probability $P(M)$ of choosing among the groups leading out of M. In case no group contains more than one element (at any rate no group from a M with $P(M) > 0$ then $|H| = |K|$ and $Q(K) = Q(M)$. This is also clear since there is then a one-to-one correspondence between the keys and messages for any given E.

From the first equation of the theorem we may conclude that $Q(K) = |K|$ in case $|M| = |E|$. This latter occurs in particular if all M's are equally likely and all E's equally likely and there are the same number of each. It is easy to see that this is the case with a language in which every letter is equally likely and independent, and when almost any of the simple ciphers are used.

If we have a product system $S = T R$, it is to be expected that the second enciphering process does not decrease the equivocation of message and this is actually true as can be shown by the methods used above. If T and R commute either may be considered as being the first and hence in this case the equivocation with S is not less than the maximum for the two systems R and T. Simple examples show that this does not hold necessarily if R and T do not commute.

Theorem 14: The equivocation in message of a product system $S = T R$ is not less than that when only R is used. If $T R = R T$ it is not less than the maximum of those for R and T alone.

If we have a product of several systems R S T U, we can of course extend this, to say that the equivocation of R S T U is not less than that of S T U, which is not less than that for T U, etc.

There is no similar theorem for the inner product since for example if T and R are inverse processes their inner product is the identity and the resulting equivocation zero.

Suppose we have a system T which can be written as a weighted sum of several systems R, S, ..., U

$$T = p_1 R + p_2 S + \dots + p_m U \quad \sum p_i = 1$$

and that systems R, S, ..., U have equivocation characteristics Q_1, Q_2, \dots, Q_m .

Theorem 15: The equivocation Q of a weighted sum of systems is bounded by the inequalities

$$\sum p_i Q_i \leq Q \leq \sum p_i Q_i - \sum p_i \log p_i$$

These are best limits possible. The Q 's may refer either to key or to message.

The upper limit is achieved, for example, in strongly ideal systems (to be described later) where the decomposition is into the simple transformations of the system. The lower limit is achieved if all the systems R, S, ..., U go to completely different cryptogram spaces. This theorem is also proved by the general inequalities governing equivocation,

$$Q_A(B) \leq Q(B) \leq Q(A) + Q_A(B).$$

We identify A with the particular system being used and B with the key or message.

There is a similar theorem for weighted sums of languages.

Theorem 16: Suppose a system can be applied to languages L_1, L_2, \dots, L_m and has equivocation characteristics Q_1, Q_2, \dots, Q_m . When applied to the weighted sum $\sum p_i L_i$, the equivocation Q is bounded by

$$\sum p_i Q_i \leq Q \leq \sum p_i Q_i - \sum p_i \log p_i$$

These limits are the best possible and the equivocations in question can be either for key or message.

The proof here is essentially the same as for the preceding case.

An important consequence of the result

$$Q(K) = |K| + |M| - |E|$$

is the following.

Theorem 17: In any closed system, or any system where the total number of possible cryptograms is equal to the number of possible messages of N letters $Q(K) \geq |K| - (|M_0| - |M|) = |K|$ where $M_0 = \log H$, with H the number of possible messages of N letters. D_N is the total redundancy for N letters.

This is true since $|M_0| \geq |E|$, the equality holds only if all cryptograms are equally likely. The theorem states that in a closed system the key is determined only by the redundancy of the language - the equivocation can decrease as the redundancy comes into action and at no greater rate

Suppose we have a pure system and let the different residue classes of messages be $C_1, C_2, C_3, \dots, C_r$. The corresponding set of residue classes of cryptograms is C_1^*, C_2^*, \dots . The probability of each E in C_1^* is the same:

$$P(E) = \frac{P(C_k)}{\phi_1} \quad E \in C_1^*$$

where ϕ_1 is the number of different messages in C_1 . Thus:

$$\begin{aligned} |E| &= -\sum_i \phi_i \frac{P(C_i)}{\phi_i} \log \frac{P(C_i)}{\phi_i} \\ &= -\sum_i P(C_i) \log \frac{P(C_i)}{\phi_i} \end{aligned}$$

Substituting in our equation for Q we obtain:

Theorem 18: For a pure cipher

$$Q = |K| + |M| + \sum_i P(C_i) \log \frac{P(C_i)}{\varphi_i}$$

This result can be used to compute Q in many cases of interest.

From the analytic point of view pure ciphers have simple structure. If a cryptogram is intercepted its residue class gives the complete information obtained by the cryptogram. Within the residue class the system is perfect - each message in the class has an a posteriori probability equal to its a priori probability. For large N, beyond the unicity point there will usually only be one M in the class of reasonable probability, and the problem is to determine this M.

The theorem on equivocation of pure ciphers can be altered to show this. We have

$$\begin{aligned} \sum_i P(C_i) \log \frac{P(C_i)}{\varphi_i} &= \sum_i P(C_i) \log P(C_i) - \sum_i P(C_i) \log \frac{k}{\varphi_i} \\ &+ \sum_i P(C_i) \log k \\ &= \sum_i P(C_i) \log P(C_i) + Q_M(K) - |K| \end{aligned}$$

Hence

$$\begin{aligned} Q(K) &= |K| + |M| + \sum_i P(C_i) \log \frac{P(C_i)}{\varphi_i} \\ &= |M| + Q_M(K) + \sum_i P(C_i) \log P(C_i) \end{aligned}$$

and

$$Q(M) = |M| - [-\sum_i P(C_i) \log P(C_i)]$$

The equivocation of message is the equivocation of message the cryptogram was intercepted less the information imparted by specification of its residue class.

21. Key Appearance Characteristic

Suppose the cryptanalyst has N letters of message and N letters of the equivalent cryptogram. Then he can calculate the a posteriori probabilities of the various keys on the basis of this information, and if N is small there will remain a certain equivocation of key. For example in simple substitution, knowing 20 letters of message and cryptogram does not disclose the entire key, since only about 12 letters of the 26 will be represented. Thus there is a residual equivocation of $\log (26-12)!$, if exactly 12 letters appear. We define the mean residual key equivocation as

$$Q_M(K) = \sum_{E,M,K} P(E,M) P_{E,M}(K) \log P_{E,M}(K)$$

when $P(E,M)$ is the a priori probability of having message M and cryptogram E, and $P_{E,M}(K)$ is the conditional probability of K with E and M given.

This may be written by obvious arguments (assuming all keys equally likely)

$$Q_M(K) = \sum_{H,K} P(M,K) \log \lambda(M,K)$$

where $\lambda(M,K)$ is the number of different keys from M in para with K, that is which go to the same E as K.

For simple substitution let P_λ be the probability that a received cryptogram of N letters has λ different letters appearing in it. Then

$$Q_M(K) = \sum P_\lambda \log (26 - \lambda)!$$

Approximately

$$Q_M(K) = \sum P_\lambda (26 - \lambda) \left[\log \frac{(26-\lambda)}{e} + \log \sqrt{2\pi(26-\lambda)} \right]$$

The bracketed terms vary slowly with λ and if $P(\lambda)$ is fairly well concentrated, we may take the bracket out replacing λ by its mean value λ_1 . This gives, after recombination

$$Q_M(K) \doteq \log (26 - \lambda_1)!$$

This residual key equivocation is shown for simple substitution on English in Fig. 12. It measures how much of the key has not been used in enciphering N letters of text on the average.

Theorem 19: $Q(K) = Q(M) + Q_M(K)$

That is, the total key equivocation (when we don't know the message) is the sum of the message equivocation and the residual key equivocation; i.e., the equivocation there would be in the key if we did know the message. This follows from the fact that the key uniquely determines the message and properties 4 and 5 in Section 1.

22. Equivocation for Simple Substitution on an Independent Letter Language

We will now calculate the mean equivocation in key or message when simple substitution is applied to a two letter language, probabilities p and q for 0 and 1, with successive letters independent. We have

$$Q_M = Q_K = -\sum P_E P_E(K) \log P_E(K)$$

The probability that E contains exactly s 0's in a particular permutation is

$$P_{E_s} = \frac{1}{2} (p^s q^{N-s} + q^s p^{N-s})$$

and the a posteriori probabilities of the identity and inverting substitutions are respectively

$$P_E(0) = \frac{p^s q^{N-s}}{(p^s q^{N-s} + q^s p^{N-s})} \quad P_E(1) = \frac{p^{N-s} q^s}{(p^s q^{N-s} + q^s p^{N-s})}$$

There are $\binom{N}{s}$ terms for each s and hence

$$Q(N) = -\sum_s \binom{N}{s} p^s q^{N-s} \log \frac{p^s q^{N-s}}{(p^s q^{N-s} + q^s p^{N-s})}$$

This may be written

$$\begin{aligned} Q(N) &= -\sum \binom{N}{s} p^s q^{N-s} [s \log p + (N-s) \log q] \\ &\quad - \log (p^s q^{N-s} q^s p^{N-s}) \\ &= -N [p \log p + q \log q] + \sum \binom{N}{s} p^s q^{N-s} \log (p^s q^{N-s} q^s) \\ &= NR + \frac{1}{2} \sum \binom{N}{s} (p^s q^{N-s} + q^s p^{N-s}) \log (p^s q^{N-s} + q^s p^{N-s}) \end{aligned}$$

For $p = 1/3$, $q = 2/3$, and for $p = 1/8$, $q = 7/8$, Q has been calculated and is shown in Fig. 13.

Now assume the language contains r different letters chosen independently and with probabilities p_1, p_2, \dots, p_r . By approximately the same argument we have

$$Q(N) = -\sum_{s_1 \dots s_r} \binom{N}{s_1 \dots s_r} p_1^{s_1} p_2^{s_2} \dots p_r^{s_r} \log \frac{p_1^{s_1} \dots p_r^{s_r}}{\sum_p p_\alpha^{s_1} \dots p_\eta^{s_r}}$$

where $\sum s_i = N$ and \sum is over all permutations of $1, 2, \dots$ for α, \dots, η .

Hence, by obvious transformations

$$Q(N) = NR + \frac{1}{r!} \sum_{s_1 \dots s_r} \binom{N}{s_1 \dots s_r} \sum_p p_\alpha^{s_1} \dots p_\eta^{s_r} \log \sum_p p_\alpha^{s_1} \dots$$

where $R = -\sum p_i \log p_i$. In particular,

$$Q(0) = \frac{1}{r!} r! \log r! = \log r! = |K|$$

$$\begin{aligned} Q(1) &= R + \frac{1}{r!} r (r-1)! \log (r-1)! \\ &= R + \log (r-1)! \end{aligned}$$

This checks the evident answer for $Q(1)$ - the r symbol has equivocation R and the parts of the key not used add $\log (r-1)!$

23. The Equivocation Characteristic for a "Random" Closed Cipher.

In the preceding section we have calculated the equivocation characteristic for a simple substitution applied to an independent letter language. This is about the simplest type of cipher and the simplest language structure possible yet already the formulas are so involved as to be nearly useless. What are we to do with cases of practical interest say the involved transformations of a fractional transposition system applied to English with its extremely complex statistical structure? This complexity itself suggests the method of approach. Sufficiently complicated problems can frequently be solved statistically. In order to do this we define the notion of a "random" cipher.

We suppose that the possible messages of length N can be divided into two groups, one group of high and fair uniform probability, while the total probability in the second group is small. This is usually possible in information theory if the messages have any reasonable length. If the total number of messages be

$$H = 2^{R_0 N}$$

where R is the maximum rate and N the number of letters. high probability group will contain about

$$S = 2^{RN}$$

where R is the statistical rate.

The deciphering operation defines a function $M = f(E)$ which can be thought of as a series of lines, k for each E going back to various M 's. By a random cipher we will mean one in which all keys are equally likely and the k lines from any E go back to random M 's. The equivocation in key is given by

$$Q(K) = \sum P(E) P_E(K) \log P_E(K)$$

The probability of exactly m lines going back to the high probability group is

$$\frac{\binom{k}{m} \left(\frac{S}{H}\right)^m \left(1 - \frac{S}{H}\right)^{k-m}}{\binom{k}{m} \left(\frac{S}{H}\right)^m \left(1 - \frac{S}{H}\right)^{k-m}}$$

If a cryptogram with m lines going to high probability messages is intercepted, the equivocation is $\log m$. The probability of intercepting such a cryptogram is easily seen to be $\frac{mH}{Sk}$.

Hence the mean equivocation is

$$Q = \frac{H}{Sk} \sum_{m=1}^k \binom{k}{m} \left(\frac{S}{H}\right)^m \left(1 - \frac{S}{H}\right)^{k-m} m \log m$$

We wish to find an approximation to this for large k . If the expected value of m , namely $\bar{m} = \frac{S}{H} k$ is $\gg 1$, the variation of $\log m$ over the range where the binomial distribution assumes large values will be small and we can replace $\log m$ by $\log \bar{m}$. This then comes out of the summation leaving the expected m . Hence in this condition

$$\begin{aligned} Q &= \log \frac{S}{H} k \\ &= \log S - \log H + \log k \\ &= |K| - |M| + |M_0| \\ &= |K| - N D. \end{aligned}$$

If \bar{m} is small compared to the large k , the binomial distribution can be approximated by a Poisson distribution.*

$$\frac{\binom{k}{m} p^m q^{k-m}}{\binom{k}{m} p^m q^{k-m}} = \frac{e^{-\lambda} \lambda^m}{m!} \quad \lambda = \frac{S}{H} k$$

Hence

$$\begin{aligned} Q &= \frac{1}{\lambda} e^{-\lambda} \sum_{m=2}^{\infty} \frac{\lambda^m}{m!} m \log m \\ &= e^{-\lambda} \sum_{m=1}^{\infty} \frac{\lambda^m}{m!} \log (m+1) \end{aligned}$$

*Fry, Probability and Its Engineering Uses, p.214.

When we write $(m + 1)$ for m . This may be used in the region where λ is near unity. For $\lambda \ll 1$ the only important term in the series is $m = 1$; omitting the others

$$\begin{aligned} Q &= e^{-\lambda} \lambda \log 2 \\ &= \lambda \log 2 \\ &= 2^{|K|} 2^{-ND} \log 2 \end{aligned}$$

Thus $Q(K)$ starts off at $|K|$, and decreases linearly with slope $-D$ out to the neighborhood of $N = |K|/D$. After a short transition region, Q follows an exponential with "half-life" distance $1/D$ if D is in alternatives per letter. If D is in digits per letter $1/D$ is the distance for a decrease by a factor of 10. The behavior is shown in Fig. 14 with the approximating curves.

By a similar argument given in the appendix, the equivocation of message can be calculated. It is

$$Q(M) = |M_0| = R_0 N \text{ for } R_0 N \ll Q(K) = |K| - DN$$

$$Q(M) = Q(K) \quad R_0 N \gg Q(K)$$

$$Q(M) = Q(K) - \phi(N) \quad R_0(N) \sim Q(K)$$

where $\phi(N)$ is the function of Fig. 14, with N scale reduced by a factor of $\frac{D}{R_0}$. $Q(M)$ rises linearly with slope R_0 until

this line intersects the $Q(K)$ line. After a rounded transition it follows $Q(K)$ down.

Most ciphers have an equivocation characteristic of this general type, approaching zero rather sharply. We will call the number of letters required for near unicity solution the unicity distance.

24. Application to Standard Ciphers.

The characteristic derived for the random cipher may be expected to apply approximately in many cases, providing some precautions are taken and certain corrections are made. The main points to be observed are the following

1. We assumed in deriving the random characteristic that the possible decipherments of a cryptogram are a random selection from the possible messages. This is not true in actual cases, but becomes more nearly true as the complexity of the operations used in the enciphering process and the complexity of the language structure increase. The more complicated the type of cipher, the more it should follow the random characteristic. In the case of

a transposition cipher it is clear that letter frequencies are preserved. This means that the possible decipherments are chosen from a more limited group - not the entire message space - and the formula should be changed. In place of R_0 one uses R_1 the rate for independent letters but with the regular frequencies. This changes the redundancy from

$$D = R_0 - R \approx .707 \text{ digits/letter}$$

to

$$D' = R_1 - R \approx .538 \text{ digits/letter}$$

and the equivocation reduces more slowly. In some other cases a definite tendency toward returning the decipherments to high probability messages can be seen. If there is no clear tendency of this sort, and the system is fairly complicated, and the language a natural one (with its very complex statistical structure) - then it is reasonable to make the random cipher assumption.

2. In many cases the key does not all appear as soon as it might. For example in simple substitution one must wait for a long time to find all letters of the alphabet represented in the message and thus deduce the complete key. The message becomes unique long before this point. Obviously our random assumption falls down in such a case, since all the different keys which differ only in the letters not yet appearing lead back to the same message, and are not randomly distributed. This error is easily corrected by the use of the key appearance characteristic. One uses at a particular N , the amount of key that may be expected at that point in the formula for Q .
3. There are certain "end effects" due to the definite starting of the message which produce a discrepancy from the random characteristics. If we take a random starting point in English text the first letter (when we do not observe the preceding letters) has a possibility of being any letter with

the ordinary letter probabilities. The next letter is more completely specified since we then have digram frequencies. This decrease in choice value continues for some time. The effect of this on the curve is that the straight line part is displaced, and approached by a curve depending on how much the statistical structure of the language is spread out over adjacent letters. As a first approximation the curve can be corrected by shifting the line over to the half redundancy point - i.e., the number of letters where the language redundancy is half its final value.

If account is taken of these three effects, reasonable estimates of the equivocation characteristic and unicity point can be made. The calculation can be done graphically as indicated in Figs. 15 and 16. One draws the key appearance characteristic $|K| - Q_M(K)$ and the total redundancy curve $|M_0| - |M|$ (which is usually sufficiently well represented by the line NR). The difference between these out to the neighborhood of their intersection is Q . For the simple substitution the characteristic is shown in Fig. 17. In so far as experimental checks could be carried out they fit this curve very well. For example, the unicity point, at about 27 letters, can be shown experimentally to lie between the limits 22 and 30. With 30 letters one nearly always has a unique solution to a cryptogram of this type and with 22 it is usually easy to find a number of them.

With transposition of period d , the unicity point occurs at about $1.5 d \log d/c$. This also checks fairly well experimentally. Note that in this case Q is defined only for integral multiples of d .

With the Vigenere the unicity point will occur at about $2d + 2$ letters, and this too is about right. The Vigenere characteristic with the same key size as simple substitution will be approximately as shown in Fig. 18. The Vigenere, Playfair and Fractional cases are more likely to follow the theoretical formulas for random ciphers than simple substitution and transposition. The reason for this is that they are more complex and give better mixing characteristics to the messages on which they operate.

The mixed alphabet Vigenere (each of d alphabets mixed independently and used sequentially) has a key size,

$$|K| = d \log 26! = 26.3 d$$

and its unicity point should be at about $53 d + 2$ letters

These conclusions can also be put to a rough experimental test with the Caesar type cipher. In the particular cryptogram analyzed in Table I, section 19, the function $Q(N)$ has been calculated and is given below, together with the values for a random cipher.

<u>N</u>	<u>0</u>	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>
Q (observed)	1.41	1.25	1.00	.60	.34	0
Q (calculated)	1.41	1.25	.98	.54	.15	.03

The agreement is seen to be quite good, especially when we remember that the observed Q should actually be the average of many different cryptograms, and that D for the larger values of N is only roughly estimated.

It appears then that the random cipher analysis can be used to estimate equivocation characteristics and the unicity distance for the ordinary types of ciphers.

25. Solving Systems Using Only N-Gram Structure.

The preceding analysis can also be applied to cases where the cryptanalyst is assumed to know or use only a limited knowledge of the structure of the language. If no data about the language other than the digram frequencies is used in solving cryptograms the equivocation curves may be computed, using for the redundancy curve that obtained from D_2 alone. This curve lies below the curve for all redundancy and the unicity point will therefore be moved to a larger N . Fig. 19 shows the Q curves for simple substitution on normal English when the cryptanalyst uses only digram structures.

26. Validity of a Cryptogram Solution.

The equivocation formulas are relevant to questions which sometimes arise in cryptographic work regarding the validity of an alleged solution to a cryptogram. In the history of cryptography one finds many cryptograms, or possible cryptograms, where clever analysts have found a "solution". It involved, however, such a complex process that the material was so scanty, that the question arose as to

whether the cryptanalyst had "read a solution" into the cryptogram. See for example the Bacon-Shakespeare cipher and the "Roger Bacon" manuscript.*

In general we may say that if a proposed system and key solves a system for a length of material considered greater than the unicity distance the solution is trustworthy. If the material is of the same order or shorter than the unicity distance the solution is highly suspicious.

This effect of redundancy in gradually producing a unique solution to a cipher can be thought of in another way which is helpful. The redundancy is essentially a series of conditions on the letters of the message, which insure that it be statistically reasonable. These consistency conditions produce corresponding consistency conditions in the cryptogram. The key gives a certain amount of freedom to the cryptogram, but as more and more letters are intercepted, the consistency conditions use up the freedom allowed by the key. Eventually there is only one message and key which satisfy all the conditions and we have a unique solution. In the random cipher the consistency conditions are in a sense "orthogonal" to the "grain of the key", and have the full effect in eliminating messages and keys as rapidly as possible. This is the usual case. However, by proper design it is possible to "line up" the redundancy of the language with the "grain of the key" in such a way that the consistency conditions are automatically satisfied and Q does not approach zero. These "ideal" systems are of such a nature that the transformations T , all induce the same probabilities in the E space. Ideal characteristics are shown in Fig. 20.

27. Ideal Secrecy Systems.

We have seen that perfect secrecy requires an infinite amount of key. With a finite key size, the equivocation of key and message generally approach zero, but not necessarily so. In fact it is possible for $Q(K)$ to remain constant at its initial value $|K|$. Then, no matter how much material is intercepted, there is not a unique solution but many of comparable probability. We will define an "ideal" system as one in which $Q(K)$ and $Q(M)$ do not approach zero as $\rightarrow \infty$. A "strongly ideal" system is one in which $Q(K)$ remains constant at $|K|$.

*See Fletcher Pratt, "Secret and Urgent"

An example is a simple substitution on an artificial language in which all letter probabilities are the same and each letter independently chosen. It is clear that $Q(K) =$ and $Q(M)$ rises linearly along a line of slope R_0 until it strikes the line $Q(K)$, after which it remains constant at this value.

With natural languages it is in general possible to approximate the ideal characteristic - the unicity point can be made to occur for as large N as is desired. The complexity of the system needed usually goes up rapidly as we attempt to do this, however. It is not always possible to actually attain the ideal characteristic with any system of finite complexity.

To approximate the ideal equivocation, one may first operate on the message with a transducer which reduces to the normal form - i.e., with all redundancies removed. After this almost any simple ciphering system - substitution, transposition, Vigenère etc., is satisfactory. The more elaborate the transducer and the nearer the output is to normal form, the more closely will the secrecy system approximate the ideal characteristic. Theorem 20: A necessary and sufficient condition that T be strongly ideal is that for any two keys $T_1, \dots, T_j, \Omega_M$ is a measure preserving transformation of Ω_M into itself.

This is true since the a posteriori probability of each key is equal to its a priori probability if and only if this condition is satisfied.

28. Examples of Ideal Secrecy Systems.

Suppose our language consists of a sequence of letters all chosen independently and with equal probabilities. Then the redundancy is zero, $|M_0| = |M|$, and from Theorem 15 $Q(K) = |K|$. We obtain the result

Theorem 21: If all letters are equally likely and independent any closed cipher is strongly ideal.

The equivocation of message will rise along the key appearance characteristic $|K| - Q_M(K)$ which will usually approach $|K|$, although in some cases it does not. In the cases of N -gram substitution, transposition, Vigenère and variations, fractional, etc., we have strongly ideal systems for this simple language with $Q(M) \rightarrow |K|$ as $N \rightarrow \infty$.

If the letters are independent but are not all equally probable, the transposition cipher characteristics remain essentially the same. The asymptotic equivocations of both key and message are clearly $|K|$. In the substitution cipher they will be less. If all the letter probabilities are different, then the asymptotic equivocations of both key and message are zero. The letters can all eventually be determined by frequency count (apart from certain exceptional sequences of zero measure). Suppose now that there are 7 letters with probabilities,

$$p_1 = p_2 < p_3 < p_4 = p_5 = p_6 < p_7$$

In this case we cannot separate p_1 from p_2 or p_4, p_5 and p_6 from each other, but the different unequal probability groups can be eventually separated.

If all substitutions are a priori equally likely, there will be an asymptotic uncertainty among

$$2! \times 3!$$

equally likely (a posteriori) keys. Hence, the asymptotic Q will be

$$Q_{\infty}(M) = Q_{\infty}(K) = \log 2! \cdot 3!$$

In general it is clear that the asymptotic equivocation with a substitution where the different substitutions are equally likely is

$$Q_{\infty}(M) = Q_{\infty}(K) = \log H$$

where H is the order of the group of substitutions on the letter probabilities $p_1 \dots p_s$ which leave this set invariant.

More generally we can consider an arbitrary pure system T and a pure language L . Suppose that T operates only "locally" on the letters of M in the sense that the n th letter of cryptogram depends only on n and a certain finite number of the letters of M in the neighborhood of the n th one:

$$e_n = f(K, n, m_n, m_{n-1}, \dots, m_{n-p}).$$

Then we can show that there is a certain subgroup of the transformations $T_1^{-1}T_j$, which are probability preserving in the language L. In the limiting cases these would consist of the identity or of the whole group $T_1^{-1}T_j$.

Theorem 22: Under these conditions the asymptotic equivocation of key is the logarithm of the order of this subgroup of measure preserving transformations.

An ideal secrecy system suffers from a number of disadvantages.

1. The system must be closely matched to the language. This requires an extensive study of the structure of the language by the designer. Also a change of statistical structure or a selection from the set of possible messages as in the case of probable words (words expected in this particular cryptogram) renders the system vulnerable to analysis.
2. The structure of natural languages is extremely complicated, and this reflects in a complexity of the transformations required to reduce them to the normal form. Thus any machine to perform this operation must necessarily be quite involved, at least in the direction of information storage, since a "dictionary" of magnitude greater than that of an ordinary dictionary is to be expected.
3. In general, reduction of a natural language to a normal form introduces a bad propagation of error characteristic. Error in transmission of a single letter produces a region of changes near it of size comparable to the length of statistical effects in the original language.

29. Multiple Substitute Ideal Systems.

There is another way of obtaining ideal or nearly ideal characteristics using multi-valued secrecy systems. Suppose our language contains only three letters with probabilities $1/8$, $3/8$ and $4/8$, and that successive letters

in a message are chosen independently. Let there be 1 substitute for the first letter, 3 for the second and 4 for the third, and choose at random among the possible substitutes for a letter. It is clear that this system is ideal. If the different probabilities are incommensurable, we cannot exactly achieve the ideal behavior, but can approximate it, by using enough substitutes, as closely as desired.

If the language is more complex, with transition probabilities, this general method can still be used, but it becomes more involved. Suppose the choice of a letter depends only on the two preceding letters, not on any more remote part of the message. The transition probabilities $p_{ij}(k)$ completely describe the statistical structure of the language. We supply substitutes for k when it follows i, j proportion to $p_{ij}(k)$. Of all our m substitutes $mp_{ij}(k)$ represent k after the pair i, j . As before one chooses from the possible substitutes for a letter at random. The cryptogram will then be a random sequence of the m substitute letters

As an example, suppose the $p_{ij}(j)$ are the only statistics of the language and the values are given by

$i \backslash j$	1	2	3
1	.1	.3	.6
2	.2	.5	.3
3	.9	.1	0

With 10 substitutes 0, 1, 2, ..., 9 we construct a substitute table assigning substitutes (chosen randomly) in proportion to the frequencies. The following is a typical key.

$i \backslash j$	1	2	3
1	7	0, 5, 6	1, 2, 3, 4, 8, 9
2	3, 9	1, 2, 5, 6, 7	0, 4, 8
3	0, 1, 2, 3, 5, 6, 7, 8, 9	4	

If a 3 follows a 2 in the message we substitute one of 0, for it, the choice being random. A second table must be supplied for the first letter of the message, corresponding to unconditional probabilities of the three letters.

Although of theoretical interest it is doubtful whether such systems would be of much use practically because of their complexity and message expansion in ordinary case. However, the first approximation to such systems, matching letter frequencies, has been used in ciphers and is standard practice in codes (where one matches word frequencies).

30. Equivocation Rate.

We now return briefly to cases where the key is not finite, but is supplied constantly, as in the Vernam system and the running key cipher. In such cases we may define equivocation "rates". One considers the equivocation $Q(N)$ of the message when N letters have been intercepted. The equivocation rate for the message is defined as the limit (assuming it exists):

$$\lim_{N \rightarrow \infty} \frac{Q(N)}{N} = Q'$$

The rate for equivocation of key would be defined similarly using the equivocation in the part of the key that has been used only, but of course these two are the same. There are results for these parameters analogous to those obtained with finite key cases. Let R' be the mean rate of using key.

Theorem 23:

$$Q' \leq R'$$

In case the equality holds we have the analogue of ideal systems where the complete information of the key goes into equivocation. If $R' > R$ the rate of the message source, we can obtain perfect secrecy - in fact we may define perfect secrecy as the case in which $Q' = R$.

In the random case we have the analogous result

$$Q' = R' - D,$$

31. Further Remarks on Equivocation and Redundancy.

We have taken the redundancy of "normal English" to be about .7 digits per letter or 50% of R_0 . This is on

the assumption that word divisions were omitted. It is an approximate figure based on statistical structure of the order of lengths of perhaps 8 letters, and assumes the text to be of an ordinary type, such as newspaper writing, literary work, etc. Various methods of calculating redundancy have been devised and will be described in the memorandum on information mentioned in the introduction. We may note here two methods of roughly estimating this number which are of cryptographic interest.

A running key cipher is a Vernam type system where in place of a random sequence of letters the key is a meaningful text. Now it is known that running key ciphers can usually be solved uniquely. This shows that English can be reduced by a factor of two to one and implies a redundancy of at least 50%. This figure cannot be reduced very much, however, for a number of reasons, unless long range "meaning" structure of English is considered.

The running key cipher can be easily improved to lead to ciphering systems which could not be solved without the key. If one uses in place of one English text, about 4 different texts as key, adding them all to the message, a sufficient amount of key has been introduced to produce a high positive equivocation rate. Another method would be to use say every 10th letter of the text as key. The intermediate letters are omitted and cannot be used at any other point of the message. This has the same effect, since the mean rate for these spaced letters must be over .8 R_0 .

These methods might be useful for spies or diplomats who could use books or magazines for the key source.

A second way of showing the high redundancy of English is to delete all vowels from a passage. In general it is possible to fill them in again uniquely and recover the original, without knowing it in advance. As the vowels constitute about 40% of the text this puts a limit on the redundancy. Actually there is considerable redundancy left the various letter and digram frequencies being far from uniform.

This suggests a simple way of greatly improving almost any simple ciphering system. First delete all vowels or as much of the message as possible without running the risk of multiple solutions, and then encipher the residue. Since this reduces the redundancy by a factor of perhaps 3 or 4 to 1, the unicity point will be moved out by this

factor. This is one way of approaching ideal systems - using the decipherer's knowledge of English as part of the deciphering system.

Two extremes of redundancy in English prose are represented by Basic English and Joyce's "Finnegans Wake". The basic English vocabulary consists of only 850 words, and a rough estimate puts the redundancy at about 70%. A cipher applied to this sort of text would rapidly approach unicity. Joyce, on the other hand, would be relatively easy to encipher. The small redundancy is disclosed by the difficulty in filling in correctly even a single missing letter from "Finnegans Wake". What the numerical value is, would be difficult to determine; it varies widely throughout the book.

The mathematical extremes of redundancy, 0 and 100 can be constructed in artificial languages. In the first we have e.g., a single possible message. $Q(M) = 0$ identically and $Q(K)$ in the random cipher case declines as rapidly as possible i.e., as rapidly as one sends information on the system. In the other extreme all letter sequences are equally likely, and any closed ciphering system is ideal.

We may refer here to a memorandum by Nyquist (Enciphering-Effect of Redundancy in Language, May 30, 1944 in which some questions of the type we are considering here are discussed.

32. Distribution of Equivocation.

A more complete description of a secrecy system applied to a language than is afforded by the equivocation characteristics can be found by giving the distribution of equivocation. For N intercepted letters we consider the fraction of cryptograms for which Q (for these particular E 's, not the mean Q) lies between certain limits. This gives a density distribution function

$$P(Q, N) \cdot dQ$$

for the probability that for N letters Q lies between the limits Q and $Q + dQ$. The mean equivocation we have previously studied is the mean of this distribution

$$\int P(Q, N) \cdot Q \cdot dQ.$$

The function $P(Q, N)$ can be thought of as plotted along a third dimension, normal to the paper, on the Q, N plane. If the language is pure, with a small influence range (compared to K) and the cipher is pure the function $P(Q, N)$ will

usually be a ridge in this plane whose highest point follows approximately the mean Q , at least until near the unicity point. In this case, or when the conditions are nearly verified, the mean Q curve gives a reasonably complete picture of the system.

On the other hand, if the language is not pure, but made up of a set of pure components

$$L = \sum p_i L_i$$

having different equivocation curves with the system, say Q_1, Q_2, \dots, Q then the total Q distribution will usually be made up of a series of ridges. There will be one for each i weighted in accordance with its p_i . The mean equivocation characteristic will be a line somewhere in the midst of these ridges and may not give a very complete picture of the situation. This is shown in Fig. 21.

A similar effect occurs if the system is not pure but made up of several systems with different Q curves. There is then a series of ridges in the $P(Q, N)$ plot, and the mean Q strikes an average which may lie between ridges and be a very improbable value of Q for a particular cryptogram. These effects are illustrated in Fig. 22.

The effect of mixing pure languages which are near to one another in statistical structure is to increase the width of the ridge. Near the unicity point this tends to raise the mean equivocation, since equivocation cannot become negative and the spreading is chiefly in the positive direction. We expect therefore, that in this region the calculations based on the random cipher should be somewhat low.

PART III

Practical Secrecy

33. The Work Characteristic

After the unicity point has been passed there will usually be a unique solution to the cryptogram. The problem of isolating this single solution of high probability is the problem of cryptanalysis. In the region before the unicity point we may say that the problem of cryptanalysis is that of isolating all the possible solutions of high probability (compared to the remainder) and determining their various probabilities.

Although it is always possible in principle to determine these solutions (by trial of each possible key for example) different enciphering systems show a wide variation in the amount of work required. The average amount of work to determine a key for a cryptogram of N letters $W(N)$ measured say in man-hours may be called the work characteristic of the system. This average is taken over all messages and all keys with their appropriate probabilities.

For a simple substitution on English the work and equivocation characteristics would be somewhat as shown in Fig. 23. The dotted portion of the curve is where there are numerous possible solutions and these must all be determined. In the solid portion after the unicity point only one solution exists in general, but if only the minimum necessary data are given a great deal of work must be done to isolate it. As more material is used the work rapidly decreases toward some asymptotic value - where the additional data no longer reduce the labor.

This is the work characteristic for the key. It is clear that after the unicity point this function can never increase. There is also a work characteristic for the message (the average amount of work to determine the message (or all reasonable messages)). This will, in ordinary cases, be below or at any rate not far above the work characteristic for the key, out to fairly large N , since generally if the key is determined it is easy to find M by the deciphering transformation. For very large N , however, this function will increase due merely to the labor of deciphering the large amount of intercepted material.

Essentially the behavior shown in Fig. 23 can be expected with any type of secrecy system where the equivocation approaches zero. The scale of man hours required, however, will differ greatly with different types of ciphers, even when the Q curves are about the same. A Vigenere or compound Vigenere, for example, with the same key size would have a much better (i.e., much higher) work characteristic. A good practical secrecy system is one in which the $W(N)$ curve remains sufficiently high out to the number of letters one expects to transmit with the key, to prevent the enemy from actually carrying out the solution, or to delay it to such an extent that the information is obsolete.

We will consider in the following sections ways of keeping the function $W(N)$ large, even though Q may be practically zero. This is essentially a "max min" type of problem as is always the case when we have a battle of wits.* In designing a good cipher we must maximize the minimum amount of work the enemy must do to break it. It is not enough merely to be sure none of the standard methods of cryptanalysis work - we must show that no method whatever will break the system easily. This, in fact, has been the weakness of many systems they were designed to resist all the known methods of solution but had a structure leading to a new method which applied to them. In the history of cryptography there have been many ciphers which were at first thought unbreakable but later disclosed weaknesses of their own.

The problem of good cipher design is essentially one of finding difficult problems, subject to certain other conditions. This is a rather unusual job for the mathematician, who ordinarily is seeking the simple and easily soluble problem in a field.

How can we ever be sure that a system which is not ideal and therefore has a unique solution for sufficiently large N will require a large amount of work to break with every method of analysis? There are two approaches to this problem

 *See von Neumann and Morgenstern, "Theory of Games". The situation between the cipher designer and cryptanalyst can be thought of as a "game" of a very single structure; a zero-sum two person game with complete information, and just two "moves". The designer chooses a system for his "move". Then the cryptanalyst is informed of this choice and chooses a method of analysis. The "value" of the play is the average work required to break a cryptogram in the system by the method chosen.

- (1) We can study the possible methods of solution available to the cryptanalyst and attempt to describe them in sufficient general terms to cover any methods he might use. We then construct our system to resist this "general" method of solution.
- (2) We may construct our ciphers in such a way that breaking it is equivalent to (or requires at some point in the process) the solution of some problem known to be laborious. Thus, if we could show that solving a system requires at least as much work as solving a system of simultaneous equations in a large number of unknown, of a complex type, then we will have a lower bound of sorts for the work characteristic.

The next three sections are aimed at these general problems. It is difficult to define the pertinent ideas involved with sufficient precision to obtain results in the form of mathematical theorems, but it is believed that the conclusions in the form of general principles, are correct.

34. Generalities on the Solution of Cryptograms

After the unicity distance has been exceeded in intercepted material, any system can be solved in principle by merely trying each possible key until the unique solution is obtained i.e., a deciphered message which "makes sense" in \mathcal{L} . A simple calculation shows that this method of solution (which we may call complete trial and error) is totally impractical except when the key is absurdly small.

Suppose, for example, we have a key of $26!$ possibilities or about 26.3 digits, the same size as in simple substitution English. This is, by any significant measure, a small key. It can be written on a small slip of paper, or memorized in a few minutes. It could be registered on 27 switches each having ten positions or on 88 two position switches.

Suppose further, to give the cryptanalyst every possible advantage, that he constructs a electronic device to try keys at the rate of one each microsecond (perhaps automatically selecting from the results by a χ^2 test for statistical significance). He may expect to reach the right key about half way through, and after an elapsed time of about

$$\frac{2 \times 10^{26}}{2 \times 60^2 \times 24 \times 365 \times 10^6} = 3 \times 10^{12} \text{ years}$$

In other words, even with a small key complete trial and error will never be used in solving cryptograms, except in the trivial case where the key is extremely small, e.g., the

caeser with only 26 possibilities, or 1.4 digits. The trial and error which is used so commonly in cryptography is of a different sort, or is augmented by other means. If one has a secrecy system which required complete trial and error it would be extremely safe. Such a system would result, it appears, in the original messages, all say of 1000 letters, were a random selection of 2 RN from the set of all 2 RoN sequences of 1000 letters. If any of the simple ciphers were applied to the messages it seems that little improvement over complete trial and error would be possible.

The methods actually used often involve a great deal of trial and error, but in a different way. First, the trials progress from more probable to less probable hypotheses, and second, each trial disposes of a large group of keys, not a single one. Thus the key space may be divided into say 10 subsets, each containing about the same number of keys. By at most 10 trials one determines which subset is the correct one. This subset is then divided into several secondary subsets and the process repeated. With the same key size ($K = 26! = 2 \times 10^{26}$) we would expect about 26×5 or 130 trials as compared to 10^{26} by complete trial and error. The possibility of choosing the most likely of the subsets first for test would improve this result even more. If the divisions were into two compartments (the best way) only 90 trials would be required. Whereas complete trial and error requires trials to the order of the number of keys, this subdividing trial and error requires only trials to the order of the key size in alternatives.

This remains true even when the different keys have different probabilities. The proper procedure then to minimize the expected number of trials is to divide the key space into subsets of equiprobability. When the proper subset is determined, this is again subdivided into equiprobability subsets. If this process can be continued the number of trials expected when each division is into two subsets will be

$$h = \frac{|K|}{\log 2}$$

If each test has S possible results and each of them corresponds to the key being in one of S equiprobability subsets then

$$h = \frac{|K|}{\log S}$$

trials will be expected. The intuitive significance of these results should be noted. In the two compartment test with equiprobability, each test yields one alternative of information as to the key. If the subsets have very different probabilities as in testing a single key in complete trial and only a small amount of information is obtained from the test. This with $26!$ equiprobable keys, a test of one yields only

$$-\left[\frac{26!-1}{26!} \log \frac{26!-1}{26!} + \frac{1}{26!} \log \frac{1}{26!} \right]$$

or about 10^{-25} alternatives of information. Dividing into S equiprobability subsets maximizes the information obtained from each trial at $\log S$, and the expected number of trials is the total information to be obtained, that is the key size, divided by this amount.

The question here is similar to various coin weighing problems that have been circulated recently. A typical example is the following: It is known that one coin in 27 is counterfeit, and slightly lighter than the rest. A chemist's balance is available and the counterfeit coin is to be isolated by a series of weighings. What is the least number of weighings to do this? The correct answer is 3, obtained by first dividing the coins into three groups of 9 each. Two of these are compared on the balance. The three possible results determine the set of 9 containing the counterfeit. This set is then divided into 3 subsets of 3 each and the process continues. The set of coins corresponds to the set of keys, the counterfeit coin to the correct key, and the weighing procedure to a trial or test.

This method of solution is feasible only if the key space can be divided into a small number of subsets, with a simple method of determining to which subset the correct key belongs. Started in another way, it is possible to solve for the key bit by bit. One does not need to assume a complete key in order to apply a consistency test and determine if the assumption is justified - an assumption on a part of the key (or as to whether the key is in some large section of the key space) can be tested.

This is one of the greatest weaknesses of most ciphering systems. For example, in simple substitution, an assumption on a single letter can be checked against its frequency, variety of contact, doubles or reversals, etc. In determining a single letter the key space is reduced by 1.4 digits from the origin

26. The same effect is seen in all the elementary types of ciphers. In the Vigenère, the assumption of two or three letters of the key is easily checked by deciphering at other points with this fragment and seeing whether clear emerges. The compound Vigenère is much better from this point of view, if we assume a fairly large number of component periods, producing a repetition rate larger than will be intercepted. Here as many key letters are used in enciphering each letter as there are periods - although this is only a fraction of the entire key, at least a fair number of letters must be assumed before a consistency check can be applied.

Our first conclusion then, regarding practical small key cipher design, is that a considerable amount of key should be used in enciphering each small element of the message.

35. Statistical Methods

It is possible to solve many kinds of ciphers by statistical analysis. Consider again simple substitution. The first thing a cryptographer does with an intercepted cryptogram is to make a frequency count. If the cryptogram contains say 200 letters it is safe to assume that few, if any, letters are out of their frequency groups, this being a division into 4 sets of well defined frequency limits. The log of the number of keys within this limitation may be calculated as

$$\log 2! 9! 9! 6! = 14.28$$

and the simple frequency count thus reduces the key uncertainty by 12 digits, a tremendous gain.

In general, a statistical attack proceeds as follows. A certain statistic is measured on the intercepted cryptogram E . This statistic is such that for all reasonable M it assumes about the same value, S_K , the value depending only on the particular key K that was used. The value thus obtained serves to limit the possible keys, to those which would give values of S in the neighborhood of that observed. A statistic which does not depend on K or which varies as much with M as with K is not of value in limiting K . Thus in transposition ciphers, the frequency count of letters gives no information about K - every K leaves this statistic the same. Hence one can make no use of a frequency count in breaking transposition ciphers.

More precisely one can ascribe a "solving power" to a given statistic S . For each value of S there will be a conditional equivocation of the key $Q_S(K)$, the equivocation

when S has its particular value and that is all that is known concerning the key. The weighted mean of these values

$$\sum P(S) Q_S(K)$$

gives the mean equivocation of the key when S is known, P being the a priori probability of the particular value S . key size $|K|$ less this mean equivocation measures the "sol-power" of S .

In a strongly ideal cipher all statistics of the togram are independent of the particular key used. This is the measure preserving property of $T_j T_k^{-1}$ on the S space or $T_j^{-1} T_k$ on the M space mentioned above.

There are good and poor statistics, just as there are good and poor methods of trial and error. Indeed the trial and error testing of hypothesis is a type of statistic, and what was said above regarding the best types of trials holds generally. A good statistic for solving a system must have the following properties:

1. It must be simple to measure.
2. It must depend more on the key than on the message if it is meant to solve for the key. The variation with M should not mask its variation with K .
3. The values of the statistic that can be "resolved" in spite of the "fuzziness" produced by variation in M should divide the key space into a number of subsets of comparable probability, with the statistic specifying the one in which the correct key lies. The statistic should give us sizable information about the key, not a tiny fraction of an alternative.
4. The information it gives must be simple and usable. Thus the subsets in which the statistic locates the key must be of a simple nature in the key space.

Frequency count for simple substitution is an example of a very good statistic.

Two methods (other than recourse to ideal systems) suggest themselves for frustrating a statistical analysis. These we may call the methods of diffusion and confusion. In the method of diffusion the statistical structure of M which leads to its redundancy is "dissipated" into long range statistics - i.e., into statistical structure involving long combinations

of letters in the cryptogram. The effect here is that the must intercept a tremendous amount of material to tie down structure, since the structure is evident only in blocks of small individual probability. Furthermore even when he has sufficient material, the analytical work required is much greater since the redundancy has been diffused over a large number of individual statistics. An example of diffusion of statistics is operating on a message $M = m_1, m_2, m_3 \dots$ with a "smoothing" operation, e.g.

$$y_n = \sum_{i=1}^s m_{n+i} \text{ mod } 26$$

adding s successive letters of the message to get a letter. One can show that the redundancy of the y sequence is the same as that of the m sequence, but the structure has been dissipated. Thus the letter frequencies in y will be more nearly equal to those in m , the digram frequencies also more nearly equal etc. Indeed any reversible operation which produces one letter out of each letter in and does not have an infinite "memory" has an output with the same redundancy as the input. The statistician can never be eliminated without compression, but they can be spread out.

The method of confusion is to make the relation between the simple statistics of E and the simple description of K complex and involved one. In the case of simple substitution was easy to describe the limitation of K imposed by the letter frequencies of E . If the connection is very involved and confused the enemy can still evaluate a statistic S_1 say which limits the key to a region of the key space. This limitation, however, is to some complex region R in the space - folded over many times and he has a difficult time making use of it. A second statistic S_2 limits K still further to R_2 , hence it lies in the intersection region $R_1 R_2$, but this does not help much because it is so difficult to determine just what the intersection is.

To be more precise let us suppose the key space has certain "natural coordinates" k_1, k_2, \dots, k_p which he wishes to determine. He measures a set of statistics s_1, s_2, \dots, s_n and these are sufficient to determine the k_i . However, in the method of confusion, the equations connecting these sets of variables are involved and complex. We have, say,

$$f_1(k_1, k_2, \dots, k_p) = s_1$$

$$f_2(k_1, k_2, \dots, k_p) = s_2$$

$$f_n(k_1, k_2, \dots, k_p) = s_n$$

and all the f_i involve all the k_i . The cryptographer must solve this system simultaneously - a difficult job. In the simple (not confused) cases the functions involve only a small number of the k_i - or at least some of these do. One first solves the simpler equations, evaluating some of the k_i and substitutes these in the more complicated equations.

The conclusion here is that for a good ciphering system steps should be taken either to diffuse or confuse the redundancy (or both).

36. The Probable Word Method

One of the most powerful tools for breaking ciphers is the use of probable words. The probable words may be words or phrases expected in the particular message due to its source, or they may merely be common words or syllables which occur in any text in the language, such as the, and, tion, that, etc.

In general, the probable word method is used as follows. Assuming a probable word to be at some point in the clear, the key or a part of the key is determined. This is used to decipher other parts of the cryptogram and provide a consistency test. If the other parts come out in clear, the assumption is justified.

There are few of the classical type ciphers that use a small key and can resist long under a probable word analysis. From a consideration of this method we can frame a test of ciphers which might be called the acid test. It applies only to ciphers with a small key (less than say 50 digits), applied to natural languages, and not using the ideal method of gaining secrecy. The acid test is this: How difficult is it to determine the key or a part of the key knowing a sample of message and corresponding cryptogram? Any system in which this is easy cannot be very resistant, for the cryptanalyst can always make use of probable words, combined with trial and error, until a consistent solution is obtained.

The conditions on the size of the key make the amount of trial and error small, and the condition about ideal systems is necessary, since these automatically give consistency checks. The existence of probable words and phrases is implied by the condition of natural languages. Conversely, it seems reasonable that if the key is difficult to obtain, knowing a text and its cryptogram, then the system should be strong.

Note that this requirement by itself is not contradictory to the requirements that enciphering and deciphering be simple processes. Using functional notation we have for enciphering

$$E = f(K, M)$$

and for deciphering

$$M = g(K, E).$$

Both of these may be simple operations on their arguments without the third equation

$$K = h(M, E)$$

being simple.

We may also point out in investigating a new type of ciphering system one of the best methods of attack is to consider how the key could be determined if a sufficient amount of M and E were given.

With a small key, the work required to solve a system, given a large amount of data, may be expected to be not more than a few orders of magnitude greater than the work required to obtain the key from a small amount of data when both M and E are known.

The same principle of confusion can be (and must be) used here to create difficulties for the cryptanalyst. Given $K = m_1 m_2 \dots m_s$ and $E = e_1 e_2 \dots e_s$ the cryptanalyst can set up equations for the different key elements $k_1 k_2 \dots$ (namely the enciphering equations).

$$e_1 = f_1(m_1, m_2, \dots, m_s; k_1, \dots, k_r)$$

$$e_2 = f_2(m_1, m_2, \dots, m_s; k_1, \dots, k_r)$$

$$e_s = f_s(m_1, m_2, \dots, m_s; k_1, \dots, k_r)$$

All is known, we assume, except the k_i . Each of these equations should therefore be complex in the k_i , and involve many of them. Otherwise the enemy can solve the simple one and then the more complex ones by substitution.

From the point of view of increasing confusion, it is desirable to have the f_i involve several m_i , especially if these are not adjacent and hence less correlated. This introduces the undesirable feature of error propagation, however, for then each e_i will generally affect several m_i in deciphering, and an error will spread to all these.

We conclude that much of the key should be used in an involved manner in obtaining any cryptogram letter from the message to keep the work characteristic high. Further dependence on several uncorrelated m_i is desirable, if some propagation of error can be tolerated. We are led by all three of the arguments of these sections to consider "mixing transformations."

37. Mixing Transformations

A notion that has proven valuable in certain branches of probability theory is the concept of a "mixing transformation." Suppose we have a probability or measure space Ω , an measure preserving transformation T of the space into itself i.e., a transformation such that the measure of a transformed region TR is equal to the measure of the initial region R . The transformation is called mixing if for any function defined over the space, and any region R .

$$\lim_{n \rightarrow \infty} \int_{T^n R} f(P) dP = \int_R dP \int_{\Omega} f(P) dP.$$

This means that any initial region of the space R under successive applications of T is mixed into the entire space Ω with uniform density. In general $T^n R$ becomes a region consisting of a large number of thin filaments spread throughout the region. As n increases the filaments become finer and their density more nearly constant.

An example of a mixing transformation is shown in Fig. 21. Here measure is identified with Euclidean area. The space is the triangle, and TAP is the point λ units of distance above point P providing this does not go outside the triangle. When the top of the triangle is reached a point is transferred first to the point directly beneath, and then over to the right an irrational fraction of the base width. If this carries the point beyond the right edge

the extra distance is measured from the left edge. Successive transforms of a square region are shown in Fig. 21. For λ very large the square is turned into a uniform grating of nearly parallel thin strips covering the triangle.

A mixing transformation in this precise sense can occur only in a space with an infinite number of points, for in a finite point space the transformation must be periodic. Speaking loosely, however, we can think of a mixing transformation as one which distributes any reasonably cohesive region in the space fairly uniformly over the entire space. If the first region could be described in simple terms, the second would require very complex ones. In the case of cryptographic interest, the original region is all of a certain simple statistical structure -- after the mix the region is distributed and the structure diffused and confused.

Good mixing transformations are often formed by repeated products of two simple non-commutating operations. See for example the mixing of pastry dough discussed by Hopf.* The dough is first rolled out into a thin slab, then folded over, then rolled, and then folded again, etc.

In a good mixing transformation of a space with natural coordinates X_1, X_2, \dots, X_S the point X_i is carried by the transformation into a point X_i^1 , with

$$X_i^1 = f_i(X_1, X_2, \dots, X_S) \quad i = 1, 2, \dots, S$$

and the functions f_i are complicated, involving all the variables in a "sensitive" way. A small variation of any one, X_3 , say, changes all the X_i^1 considerably. If X_3 passes through its range of possible variation the point X_i^1 traces a long winding path around the space.

Various methods of mixing applicable to statistical sequences of the type found in natural languages can be devised. One which looks fairly good is to follow a preliminary transposition by a sequence of alternating substitutions and simple linear operations, adding adjacent letters mod 26 for example.

Thus

$$H = \text{LSLSLT}$$

where T is a transposition, L is a linear operation, and S is a substitution.

*E. Hopf, On Causality, Statistics and Probability, Journal of Math. and Physics, V.13, pp.51-102, 1934.

38. Ciphers of the Type $T_k H S_j$

Suppose that H is a good mixing transformation that can be applied to sequences of letters and that T_k and S_j any two simple families of transformations, i.e., two ciphers, which may be the same. For concreteness we may take of them as both simple substitutions.

It appears that the cipher THS will be a very good ciphering system from the standpoint of its work character. In the first place it is clear on reviewing our arguments and statistical methods that no simple statistics will give information about the key - any significant statistics derived from the ciphertext must be of a highly involved and very sensitive type - the redundancy has been both diffused and confused by the mixing. Also probable words lead to a complex system of equations involving all parts of the key (when the mix is good), which must be solved simultaneously. The bad features of such a system - propagation of errors and complexity of operations, both of which get worse as the mixing of H gets better.

It is interesting to note that if the cipher T is omitted the remaining system is similar to S and thus no stronger. The enemy merely "unmixes" the cryptogram by application of H^{-1} and then solves. If S is omitted the remaining system is much stronger than T alone if the mix is good but still not comparable to THS .

The basic principle here of simple ciphers separated by a mixing transformation can of course be extended. For example one could use

$$T_k H_1 S_j H_2 R_1$$

with two mixes and three simple ciphers. One can also simplify by using the same ciphers, and even the same keys (inner product) as well as the same mixing transformations. This might well simplify the mechanization of such systems.

The mixing transformation which separates the two (or more) appearances of the key acts as a kind of barrier to the enemy -- it is easy to carry a known element over this barrier but an unknown (the key) does not go easily.

By supplying two sets of unknowns, the key for S and the key for T , and separating them by the mixing transformation H we have "tangled" the unknowns together in a way that makes solution very difficult.

Although systems constructed on this principle would be extremely safe they possess one grave disadvantage. If the mix is good then the propagation of errors is bad. A transmission error of one letter will affect several letters on deciphering.

39. The Compound Vigenere

In the compound Vigenere several keys of length d_1, d_2, \dots, d_s are written under the message and added to it modulo 26 to obtain the cryptogram. The result is a Vigenere with key of special type, whose repetition is of period d , the least common multiple of d_1, d_2, \dots, d_s . If we have three keys of periods 2, 3, 5 the total period d is 30 and the total key size $(2+3+5) \times 1.41 = 14.1$ digits. The situation is then

$$M = m_1 m_2 m_3 m_4 m_5 m_6$$

$$K_1 = a_1 a_2 a_1 a_2 a_1 a_2$$

$$K_2 = b_1 b_2 b_3 b_1 b_2 b_3$$

$$K_3 = c_1 c_2 c_3 c_4 c_5 c_1$$

$$E = e_1 e_2 e_3 e_4 e_5 e_6$$

with

$$e_1 = m_1 + a_1 + b_1 + c_1$$

$$e_2 = m_1 + a_2 + b_1 + c_2$$

etc.

If we assume M and E known then, letting $h_1 = e_1 - m_1$:

$$a_1 + b_1 + c_1 = h_1$$

$$a_2 + b_3 + c_1 = h_5$$

$$e_2 + b_2 + c_2 = h_2$$

$$a_1 + b_1 + c_2 = h_7$$

$$a_1 + b_3 + c_3 = h_3$$

$$a_2 + b_2 + c_3 = h_8$$

$$a_2 + b_1 + c_4 = h_4$$

$$a_1 + b_3 + c_4 = h_9$$

$$a_1 + b_2 + c_5 = h_5$$

$$a_2 + b_1 + c_5 = h_{10}$$

These equations are easily solved for the key, although not as easily as in the simple Vigenère or other simple ciphers. As the number of constituent periods increases the solution becomes more involved and time consuming. In any case we have a system of simultaneous equations each involving S of the

total of $B = \sum_{i=1}^S d_i$ unknowns. The unicity point will occur at about

2B letters and if several times this amount of material is intercepted no great difficulty should be encountered in breaking the cipher, providing S is not more than say 6 or 8. With the first 9 primes as periods we have a key size of 100 letters or about 141 digits, the unicity distance is about 200 letters and the key does not repeat for 223,092,870 letters. This system, although much better than such methods as simple substitution, transposition and simple Vigenère with equivalent key size, does not utilize the available key fully in making the cryptanalyst work for the solution. The equations only involve S of the B key unknowns and these in a simple fashion. The equations easily combine and reduce to eliminate unknowns. If a large amount of material is available, compared to the unicity distance, particular sets of equations can be combined to eliminate unknowns very easily. The system possesses the important advantage, however, of not expanding errors. One incorrect letter of cryptogram produces one incorrect letter of deciphered text.

By relatively simple changes this system could be strengthened considerably. If the equations for the key elements (with M and E known) could be made into higher degree equations rather than linear ones the difficulty of solution would increase tremendously. This could easily be done in a mechanical device by successive multiplications (Mod 26) of the key letters according to some prearranged scheme,

40. Incompatibility of the Criteria for Good Systems

The five criteria for good secrecy systems given in section 12 appear to have a certain incompatibility when applied to a natural language with its complicated statistical structure. With artificial languages having a simple statistical structure it is possible to satisfy all requirements simultaneously, by means of the ideal type ciphers. In natural languages it seems that a compromise must be made and the valuations balanced against one another with a view toward the particular application.

If any one of the five criteria is dropped, the other four can be satisfied fairly well, as the following examples show.

1. If we omit the first requirement (amount of secrecy) any simple cipher such as simple substitution will do. In the extreme case of omitting this condition completely, no cipher at all is required and one sends the clear!
2. If the size of the key is not limited the Vernam system can be used.
3. If complexity of operation is not limited, various extremely complicated types of enciphering process can be used. The modified compound Vigenere described above with many different periods compounded is fairly satisfactory as an example here, although it falls down somewhat on the key size condition. Ideal systems and enciphered codes are also fair examples although not too good from the propagation of error point of view.
4. If we omit the propagation of error condition system of the type THS would be very good, although somewhat complicated.
5. If we allow large expansion of message, various systems are easily devised where the "correct" message is lost with many "incorrect" ones (misinformation). The system determines which of these is correct.

A rough argument for the incompatibility of the conditions may be given as follows.

From condition 5, secrecy systems essentially as studied in this paper must be used; i.e., no great use of etc. Perfect and ideal systems are excluded by condition 2 and by 3 and 4, respectively. The high secrecy required by must then come from a high work characteristic, not from a high equivocation characteristic. If the key is small, the system simple, and the errors do not propagate, probable work methods will generally solve the system fairly easily, since we then have a fairly simple system of equations for the key.

This reasoning is too vague to be conclusive, but the general idea seems quite reasonable. Perhaps if the various criteria could be given quantitative significance, some sort of an exchange equation could be found involving them and giving the best physically compatible sets of values. The two most difficult to measure numerically are the complexity of operations, and the complexity of statistical structure of the language.

Appendix 1

Deduction of $-\sum p_i \log p_i$

It will be shown that the measure of choice $-\sum p_i \log p_i$ is a logical consequence of three quite reasonable assumptions about the desired properties of such a measure. The three assumptions are:

(1) There exists a function $C(p_1, p_2, \dots, p_n)$ unique in the p_i , measuring the amount of "choice" when there are n possibilities with probabilities p_i .

(2) C has the property that if a given choice be broken down into two successive choices the total amount of choice is the weighted sum of the individual choices. For example, suppose the choice is from 4 possibilities A, B, C, D with probabilities .1, .2, .3, .4. This can be broken down a preliminary choice between the pair A, B and the pair C, D. Pair A, B has a total probability .1 + .2 = .3 and pair C, D probability .3 + .4 = .7. If pair A, B is chosen a second choice between A and B must be made with probabilities $\frac{.1}{.1 + .2} = \frac{1}{3}$

$\frac{.2}{.1 + .2} = \frac{2}{3}$. If pair C, D is chosen a second choice between C and D must be made with probabilities $\frac{.3}{.3 + .4} = \frac{3}{7}$ and $\frac{.4}{.3 + .4} = \frac{4}{7}$. Thus broken down we have a preliminary amount of choice $C(.3, .7)$ and of the time a secondary choice of C ($\frac{1}{3}, \frac{2}{3}$) while .7 of the time the secondary choice is C ($\frac{3}{7}, \frac{4}{7}$). Our condition requires that the total choice $C(.1, .2, .3, .4)$ be the same as the weighted sum of the different choices when decomposed, weighted in accordance with the frequency of occurrence. Thus we require in this case $C(.1, .2, .3, .4) = C(.3, .7) + .3.C(\frac{1}{3}, \frac{2}{3}) + .7.C(\frac{3}{7}, \frac{4}{7})$.

(3) If $A(n) = C(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n})$, i.e. the choice when there are n equally likely possibilities, then $A(n)$ is monotonic increasing in n .

Theorem: Under these three assumptions

$$C(p_1, p_2, \dots, p_n) = -K \sum p_i \log p_i$$

where K is a positive constant.

From condition (2) we can decompose a choice from S^m equally likely possibilities into a series of m choices each from S equally likely possibilities and obtain

$$A(S^m) = m A(S)$$

Similarly

$$A(t^n) = n A(t)$$

We can choose n arbitrarily large and find an m to satisfy

$$S^m \leq t^n < S^{m+1}$$

Thus, taking logarithms and dividing by $n \log S$,

$$\frac{m}{n} \leq \frac{\log t}{\log S} \leq \frac{m}{n} + \frac{1}{n} \quad \text{or} \quad \left| \frac{m}{n} - \frac{\log t}{\log S} \right| < \epsilon$$

where ϵ is arbitrarily small,

Now from the monotonic property of $A(n)$

$$A(S^m) \leq A(t^n) \leq A(S^{m+1})$$

$$m A(S) \leq n A(t) \leq (m+1) A(S)$$

Hence, dividing by $n A(S)$,

$$\frac{m}{n} \leq \frac{A(t)}{A(S)} \leq \frac{m}{n} + \frac{1}{n} \quad \text{or} \quad \left| \frac{m}{n} - \frac{A(t)}{A(S)} \right| < \epsilon$$

$$\left| \frac{A(t)}{A(S)} - \frac{\log t}{\log S} \right| \leq 2 \epsilon \quad A(t) = -K \log t$$

where K must be positive to satisfy (3).

Now suppose we have a choice from n possibilities with commensurable probabilities $p_1 = \frac{n_1}{\sum n_i}$ where the n_i are integers.

can break down a choice from $\sum n_i$ possibilities into a choice from n possibilities with probabilities p_1, \dots, p_n and then, if the i th was chosen, a choice from n_i with equal probabilities. Using condition 2 again, we equate the total choice from $\sum n_i$ as computed by two methods

$$K \log \sum n_i = C(p_1, \dots, p_n) + K \sum p_i \log n_i$$

Hence

$$C = K [\sum p_i \log \sum n_i - \sum p_i \log n_i]$$

$$= -K \sum p_i \log \frac{n_i}{\sum n_i} = -K \sum p_i \log p_i$$

If the p_i are incommensurable, they may be approximated by rationals and the same expression must hold by our continuity assumption. Thus the expression holds in general. The choice of coefficient K is a matter of convenience and amounts to the choice of a unit of measure.

Appendix 2Proof of Theorem 4

Select any message M_1 and group together all cryptograms that can be obtained from M_1 by an enciphering operation T_i . Let this class of cryptograms be C_1 . Group with M_1 all M_K that can be obtained from M_1 by $T_i^{-1} T_j M_1$, and call this class C_1 . The same C_1 would be obtained if we started with any other M in C_1 since

$$T_j T_i^{-1} T_i M_1 = T_j M_1$$

Similarly the same C_1 would be obtained.

Choosing an M (if any exist) not in C_1 we construct C_2 and C_2 in the same way. Thus we obtain the residue classes with properties (1) and (2). Let M_1 and M_2 be in C_1 and suppose

$$M_2 = T_2 T_1^{-1} M_1$$

If E_1 is in C_1 and can be obtained from M_1 by

$$E_1 = T_\alpha M_1 = T_\beta M_1 = \dots = T_\eta M_1,$$

then

$$E_1 = T_\alpha T_2 T_1 M_2 = T_\beta T_2^{-1} T_1 M_2 = \dots$$

$$= T_\lambda M_2 = T_\mu M_2 \dots$$

Thus each M_1 in C_1 transforms into E_1 by the same number of keys. Similarly each E_1 in C_1 is obtained from any M in C_1 by the same number of keys. It follows that this number of keys is a divisor of the total number of keys and hence we have properties (3) and (4).

Appendix 3

Equivocation of Message for Random Cipher

As before let $M_1 \dots M_S$ be high probability mes and $M_{S+1} \dots, M_H$ have zero probability. Let $P(m_1, m)$ be probability of just m_1 lines going from a particular E , s to a particular high probability M , say M_1 , with a total lines to all high probability M . Then

$$P(m_1, m) = \frac{(k)}{(m)} \frac{(m)}{(m_1)} \frac{(1)}{(H)}^{m_1} \frac{(S-1)}{(H)}^{m-m_1} \frac{(1-S)}{(H)}^{k-n}$$

The probability of intercepting an E with m lines to high bility M 's is

$$\frac{m}{Sk}$$

The $Q(M)$ expected can be thought of as contributed to by various M_1 in the high probability group. Thus M_1 contri

$$- \frac{m_1}{m} \log \frac{m_1}{m} = \frac{m_1}{m} \log \frac{m}{m_1}$$

if there are m_1 lines to M_1 and a total of m to high prob M 's. The expected Q is then

$$Q(M) = H S \sum_{m_1} P(m_1, m) \frac{m}{Sk} \frac{m_1}{m} \log \frac{m}{m_1}$$

The factor H sums over the various E_i and the S sums over different M_1 ($i = 1, \dots, S$). Hence,

$$Q(M) = \frac{H}{k} \sum P(m_1, m) m_1 [\log m - \log m_1]$$

the term

$$\sum P(m_1, m) m_1$$

summed on m_1 , gives the expected m_1 , when m lines go to i probability M_s , i.e., m/s . Hence the first term is

$$\frac{H}{ks} \sum_m m P(m) \log m = Q(K)$$

by our previous work. The second term is

$$- \frac{H}{k} \sum P(m_1, m) m_1 \log m_1$$

If the expected m_1 is $\ll 1$ this term is small since it vanishes for $m_1 = 0$ or 1. The expected m_1 is k/H . Thus beyond this point $Q(M)$ approaches closely to $Q(K)$. The point in question is where $|K| = |M_0| = R_0 N$

or

$$N = \left| \frac{K}{R_0} \right|$$

If the expected $m_1 \gg 1$ the $\log m_1$ can be taken out as $\log \bar{m}_1 = \log k/H_1$ and we have

$$\begin{aligned} & - \frac{H}{K} \log \frac{k}{H} \sum P(m_1, m) m_1 \\ & = -\log \frac{k}{H} = |M_0| - |K| \end{aligned}$$

In this region then

$$Q(M) = |M_0| - |K| + Q(K)$$

but here $Q(K) = |K| - |M_0| + |M|$, and therefore

$$Q(M) = |M| = RN$$

In the transition region \bar{m}_1 is about 1 and \bar{m} will in ordinary cases be very large. It is admissible then to replace $P(m_1, m)$ by $P(m_1)$, since this will not depend on m to any extent except for values of m of very small probability. Thus we obtain for this region

$$Q(M) = Q(K) - \frac{H}{K} \sum_{m_1=1}^K P(m_1) m_1 \log m_1$$

The sum has the same form as our expression for $Q(K)$ but with $1/H$ in place of S/H . The calculations for $Q(K)$ can be used, therefore, with only a change of the N scale by a factor of R_0/D .

Appendix 4

Key Appearance in Simple Substitution with Independent Letters

If successive letters are chosen independently and the different letters have probabilities p_1, p_2, \dots, p_S , we calculate the expected number of different letters when N letters have been intercepted. It is

$$\bar{d}(N) = S - \sum_{i=1}^S (1 - p_i)^N$$

To prove this, imagine all the possible sequences of N letters written down, each with a frequency corresponding to its probability, giving a total of say A sequences. Letter 1 does not appear in $(1 - p_1)^N A$ of these; letter 2 does not appear in $(1 - p_2)^N A$ etc. Therefore, the total number of letters missing from sequences is

$$A \sum_{i=1}^S (1 - p_i)^N$$

Dividing by A gives us by definition the expected number of missing letters from a random sequence, $\sum (1 - p_i)^N$. The number of different letters expected in a sequence is the total number of letters S minus this, giving the desired result.

If all the p_i are equal this reduces to $S - S(1 - p)^N$, an exponential approach to S . In the general case there is a series of exponentials with different time constants, corresponding to different p_i , which are added to give $\bar{d}(N)$.

With the frequencies of normal English used for p_i , we obtain the curve shown in Fig. 25, along with an exponential curve. The small discrepancy can be attributed to influences of nearby letters. In English there is less tendency to double letters than there would be if the letters were independent but with the same probabilities. For English the probability of a doubled letter is

$$\sum p_i^2 = .0315$$

while if letters were independent it would be

$$\sum p_i^2 = .0670.$$

~~CONFIDENTIAL~~

Appendix 5

A Theoretical Case Where All Invariant Statistics of E Are Independent of K.

By an invariant statistic of a sequence of letters $E = \dots m_{-2} m_{-1} m_0 m_1 m_2 \dots m_3 \dots$, we will mean a statistic which is averaged along the length of the sequence E. More precisely a statistic of the form:

$$\lim_{n \rightarrow \infty} \frac{1}{(2n+1)} \{ F(E_{-n}) + \dots + F(E_{-1}) + F(E) + F(E_1) + F(E_2) + \dots + F(E_n) \}$$

where F is any function whose argument is a possible sequence, and $E \pm n$ is the sequence E shifted N letters to the right or left. Such statistics as the relative frequency of a given letter, of a given n-gram, transition frequencies, and frequencies with which letter i is followed by letter j at a distance n are all invariant.

We will describe a system in which every invariant statistic which the cryptanalyst can construct from the (infinite) intercepted E is independent of both K and M, and thus gives no information to him. This effect and still more occurs with the ideal ciphers of course, but here it is obtained independently of the original message statistics and without any matching of the cipher to the language.

Let N be a "random" sequence of letters;

$$N = \dots n_{-2} n_{-1} n_0 n_1 n_2 \dots n_s \dots$$

this is supposedly a known sequence (to the enemy) and thus a part of the system, not of the key. Apply any simple cipher to the message and then add N letter by letter to the result (mod 26). The "sum" is the enciphered message. It is evident that any invariant statistic on E will be (with probability 1) the same as that for a random sequence. Hence it is independent of both K and M.

We need hardly add that such a system is easily broken - the enemy merely subtracts N from E and then solves the simple residual cipher, which may often be done with invariant statistics.

Appendix 6

Maximum Repetition Rate in Compound Systems for a Given Total Key Size

We consider briefly the question of how to arrange component periods in a compound Vigenère or Transposition system to obtain the longest period for a given total key size. If the component periods are P_1, P_2, \dots, P_s , it is clear that they must be coprime. Otherwise the total key, which is $\sum P_i$, could be reduced without changing the period, which is the least common multiple of the P_i , merely by deleting a factor which appears in several of the P_i from all but one. Also each P_i must be a power of a prime, for if it contains two primes, it can be divided into these parts, reducing the key and not affecting the period. The component periods are selections from the series of primes and powers of primes:

A: 2, 3, 4, 5, 7, 8, 9, 11, 13, 16, 17, 19, 23, 25, 27, ...
the selection being pairwise coprime.

It appears from empirical evidence that the best selection of component periods for a given total size S is found by the following process.

1. Determine the largest M such that $\sum_{i=1}^M p_i \leq S$ where the p_i are the primes in increasing order. This is the maximum number of periods where the periods are coprime, and is the number of periods to be used.
2. Choose from the sequence A, M elements, consecutively except for the fact that no prime is represented more than once, the M elements being as great as possible with sum $\leq S$.
3. If the sum is $< S$ move as many as possible of the elements in this block up a notch in the sequence still satisfying the conditions on the sum and coprimality.
4. Repeat 3 to either part of the original block if possible. This process eventually ends and apparently gives the proper decomposition.

For example with $S = 50$, the sum of the first primes is 41, of the first 7 is 58. Hence 6 periods will be used. We have

$$11 + 9 + 8 + 7 + 5 + 3 = 43$$

$$13 + 11 + 9 + 8 + 7 + 5 = 53$$

~~CONFIDENTIAL~~

hence we start with the block 11, 9, 8, 7, 5, 3.
The top 4 elements 11, 9, 8, 7 can be moved up a
to give

$$13 + 11 + 9 + 8 + 5 + 3 = 49$$

No further improvement seems possible. we obtain

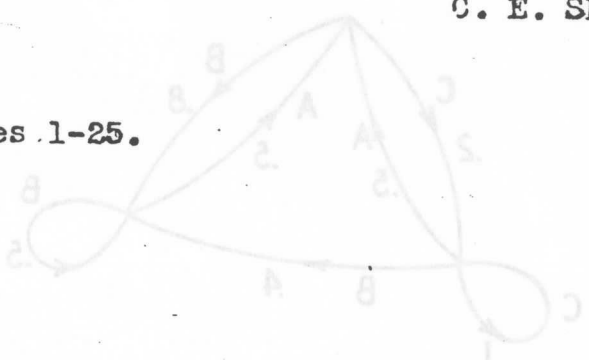
$$P = 13 \times 11 \times 9 \times 8 \times 5 \times 3 = 154,440$$

The products and sums of the first n primes are given below

n	1	2	3	4	5	6	7	8
pn	2	3	5	7	11	13	17	19
Sum	2	5	10	17	28	41	58	77
Product	2	6	30	210	2310	30030	510510	9699590
								223092

C. E. SHANNON

Att.
Figures 1-25.



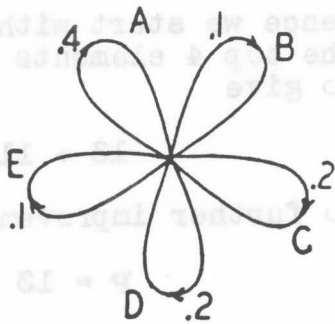


FIG. 1

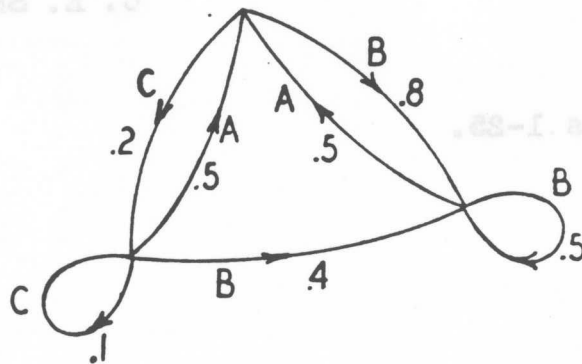


FIG. 2

~~CONFIDENTIAL~~

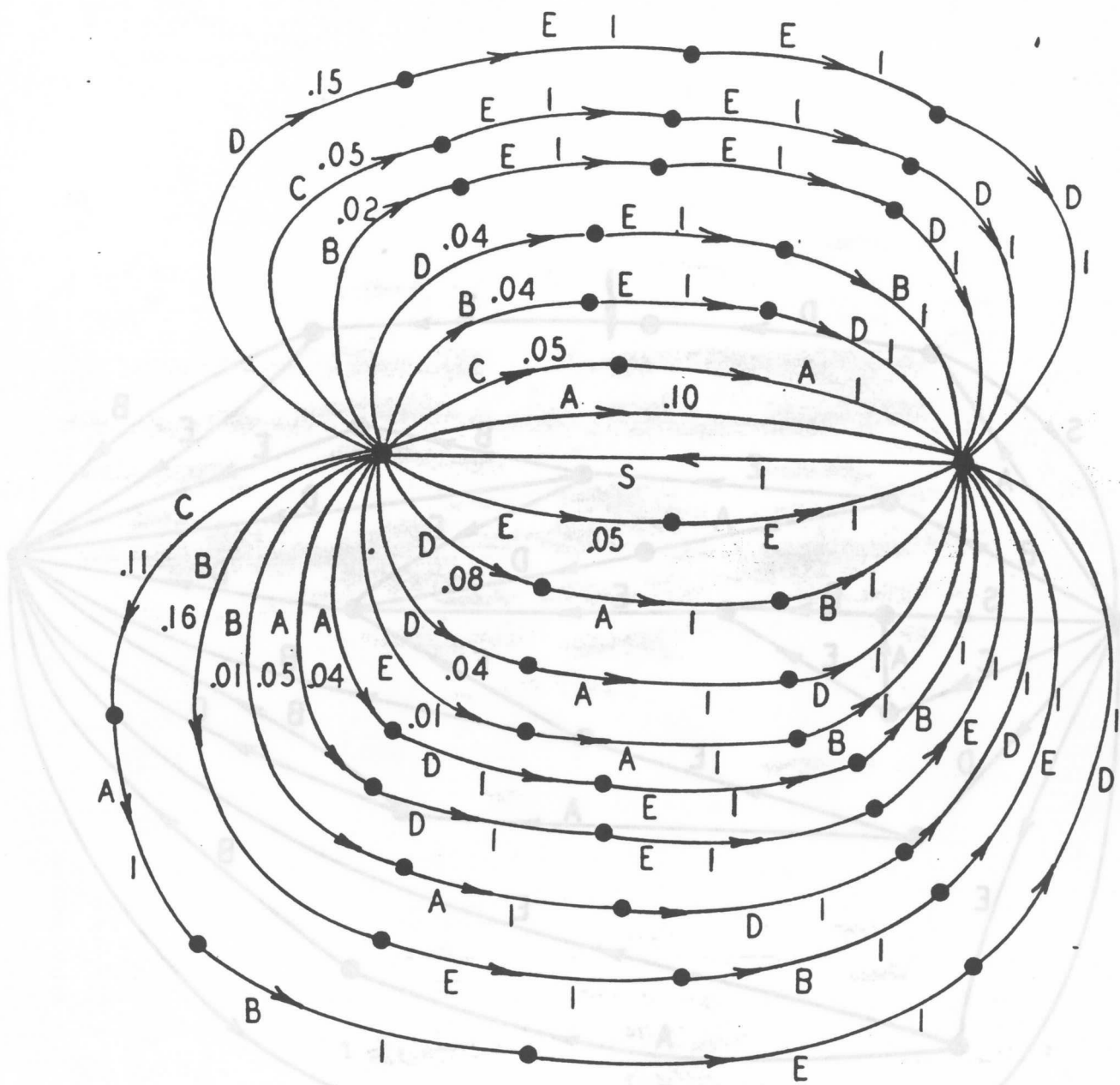


FIG. 3

~~CONFIDENTIAL~~

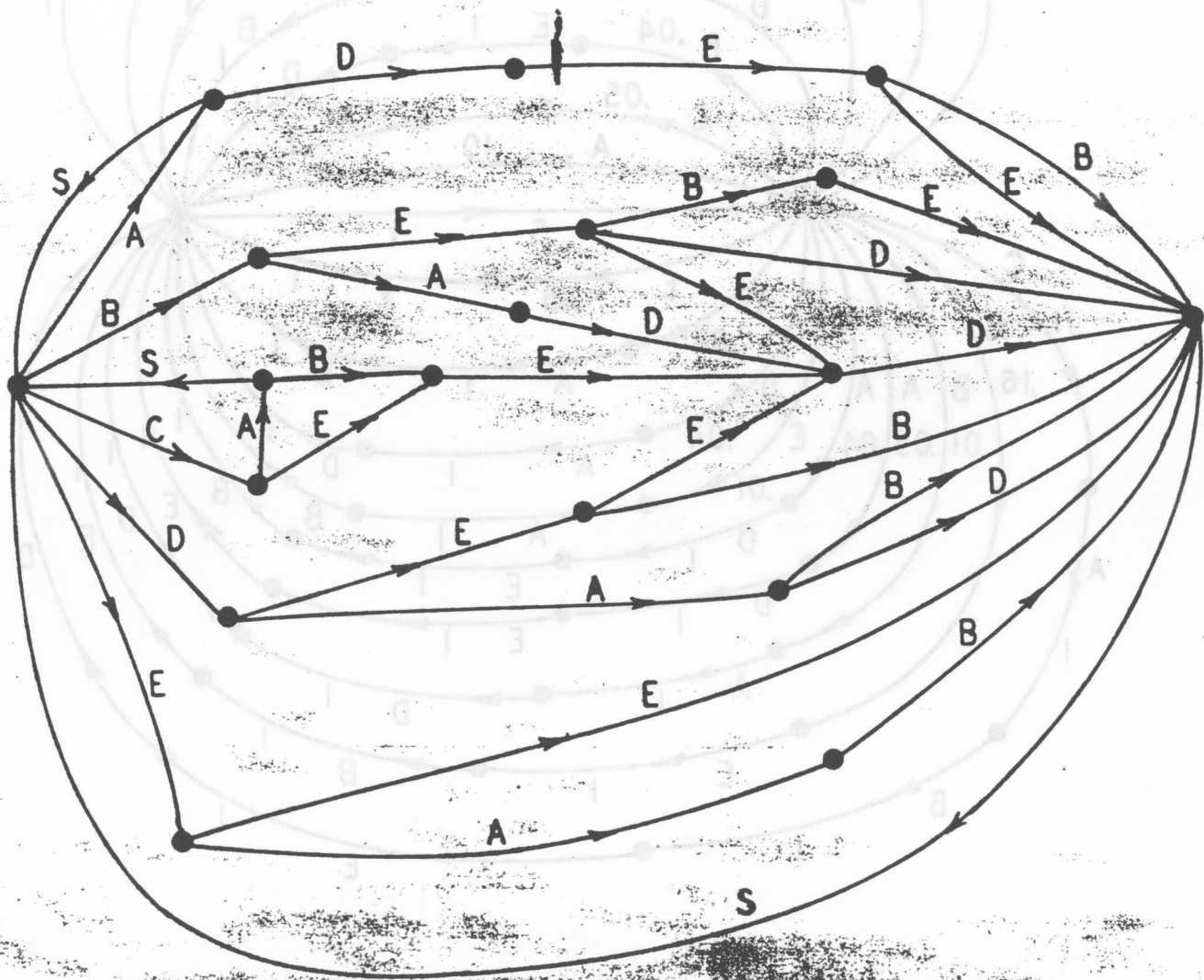


FIG. 4

~~CONFIDENTIAL~~

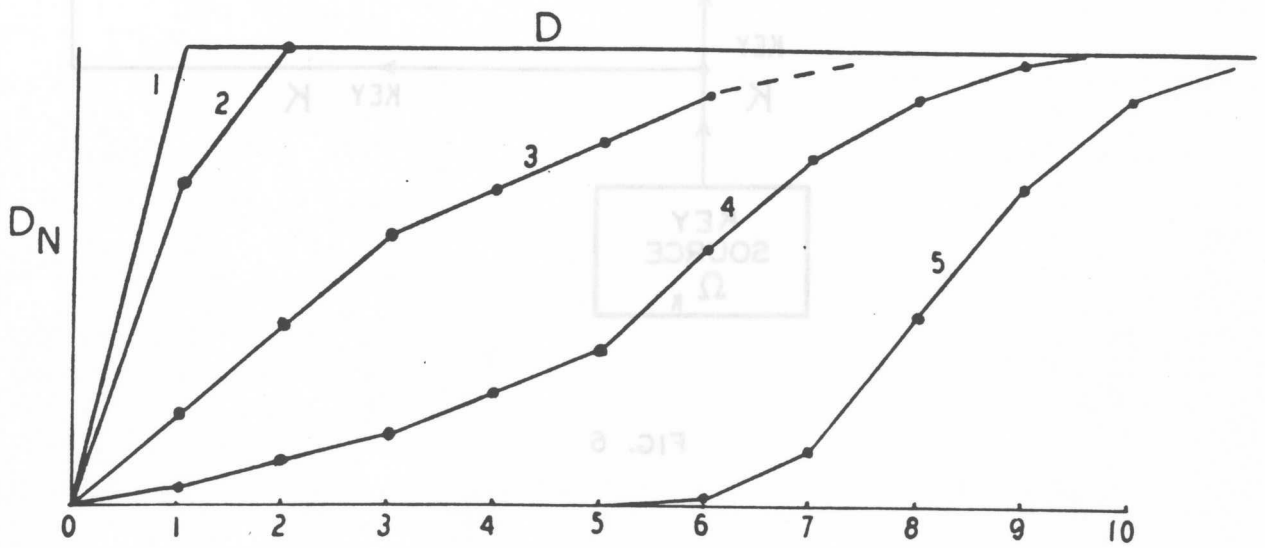


FIG. 5

~~CONFIDENTIAL~~

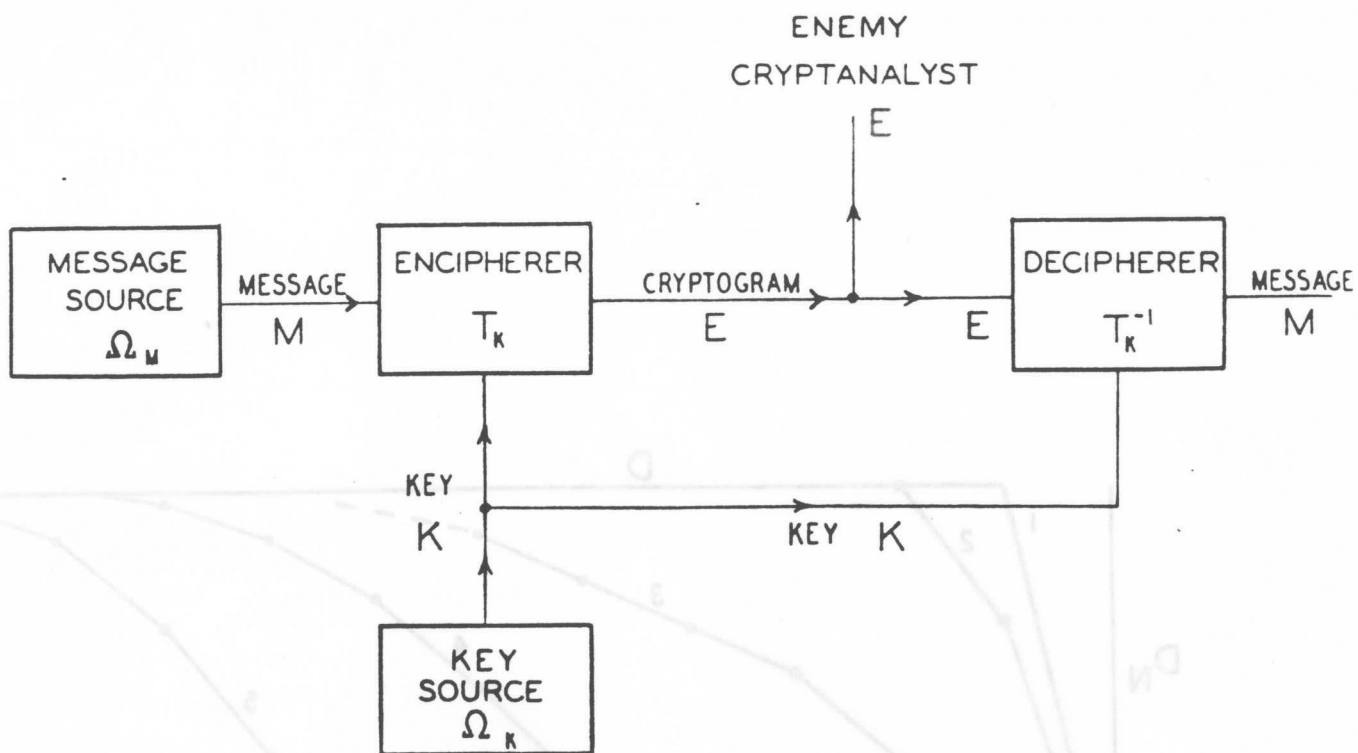


FIG. 6

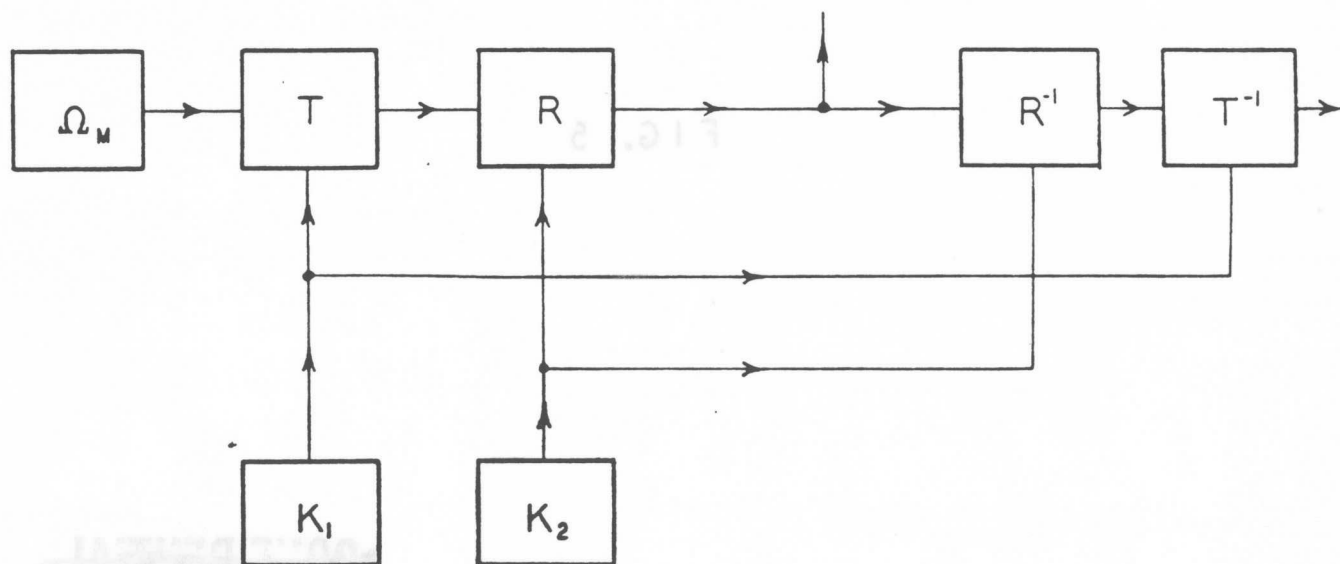


FIG. 8

~~CONFIDENTIAL~~

NOT CLOSED

CLOSED SYSTEM

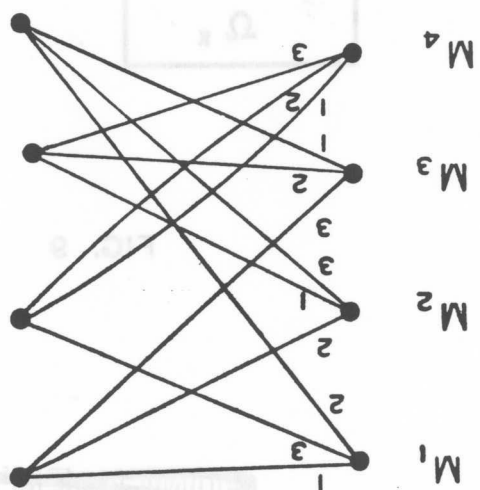
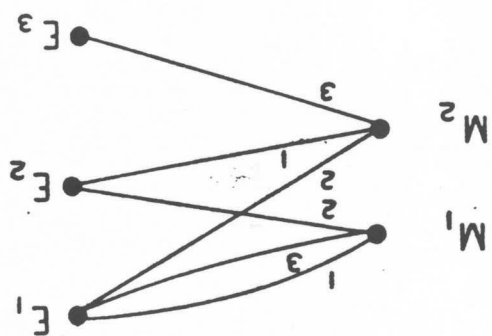
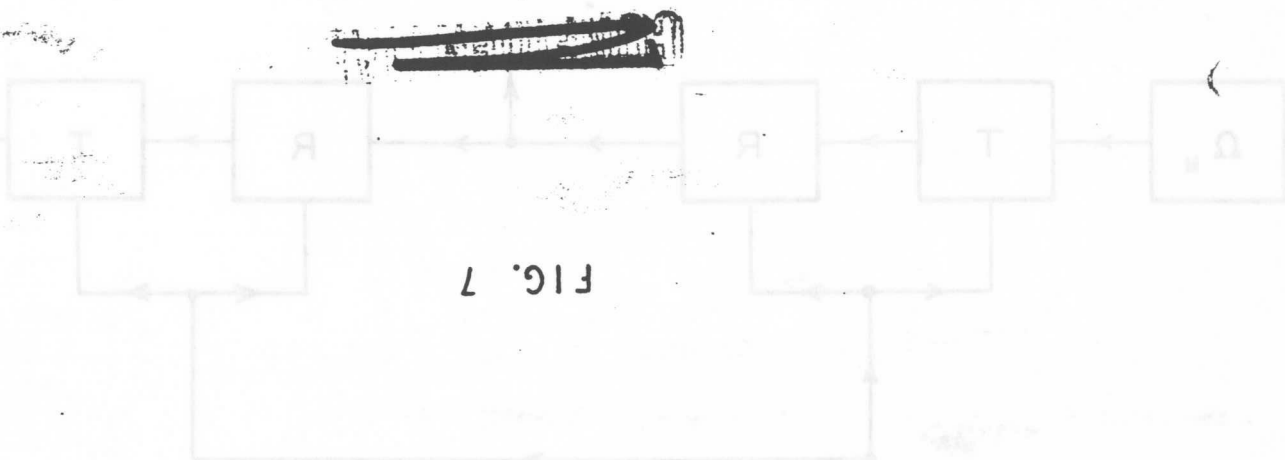


FIG. 7



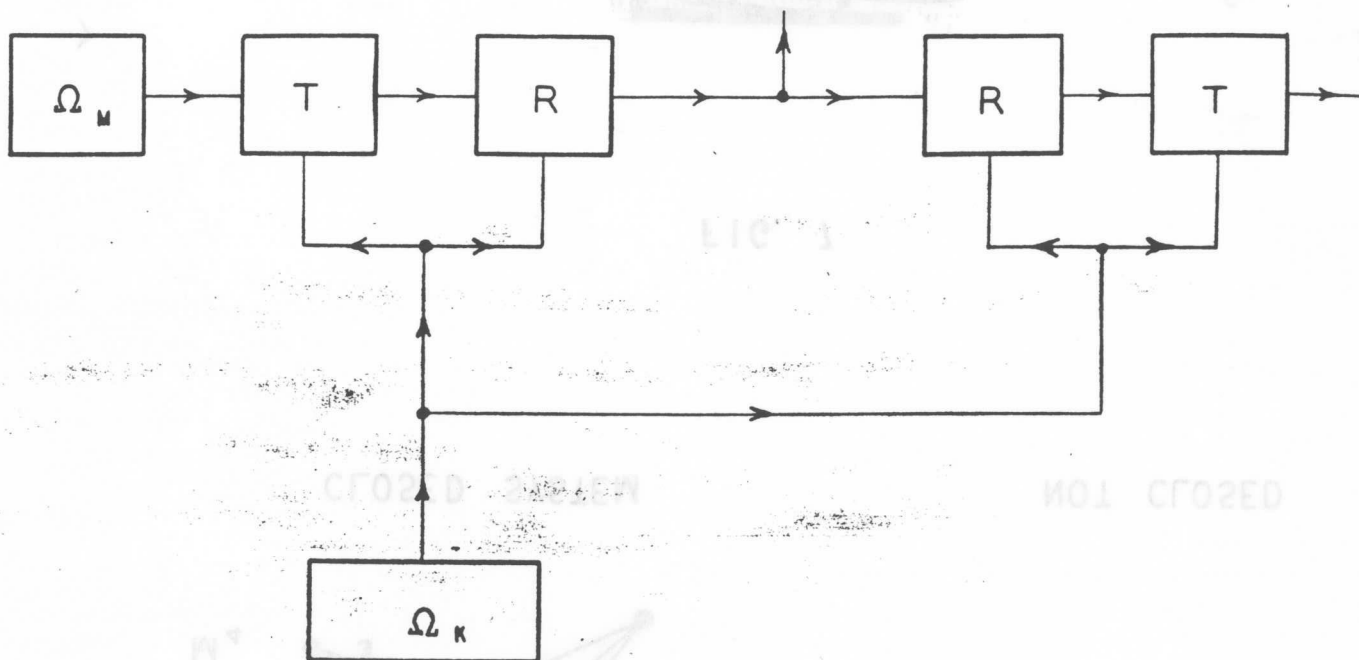
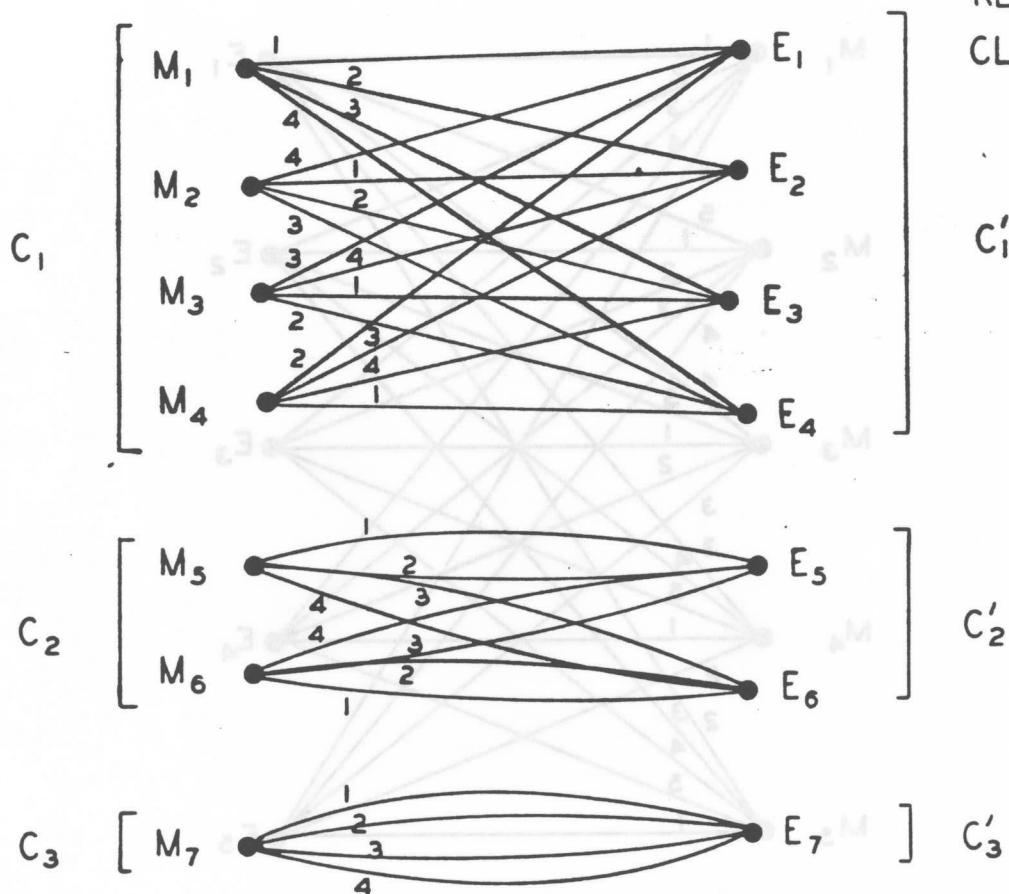


FIG. 9

~~CONFIDENTIAL~~

MESSAGE
RESIDUE
CLASSES

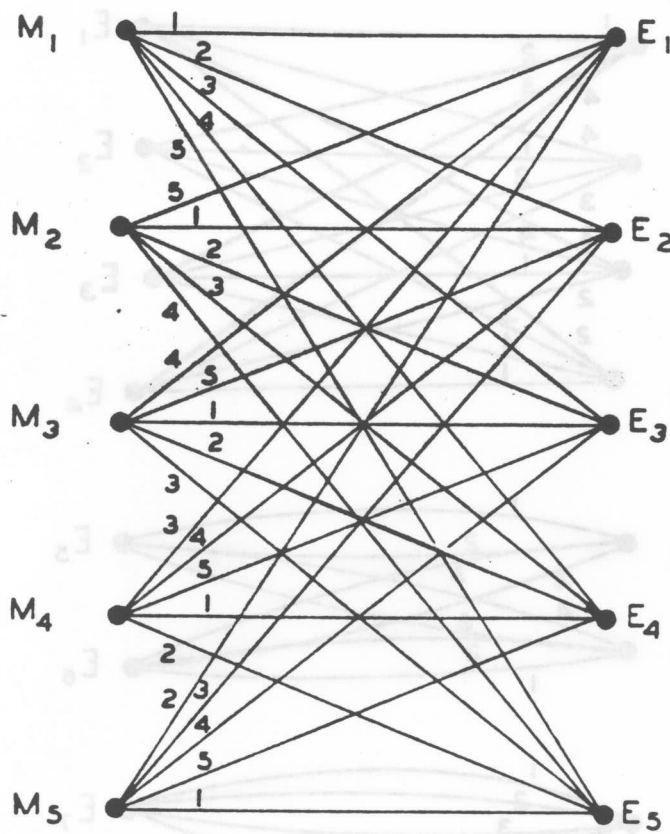
CRYPTOGRAM
RESIDUE
CLASSES



PURE SYSTEM

FIG. 10

~~CONFIDENTIAL~~



PERFECT SYSTEM

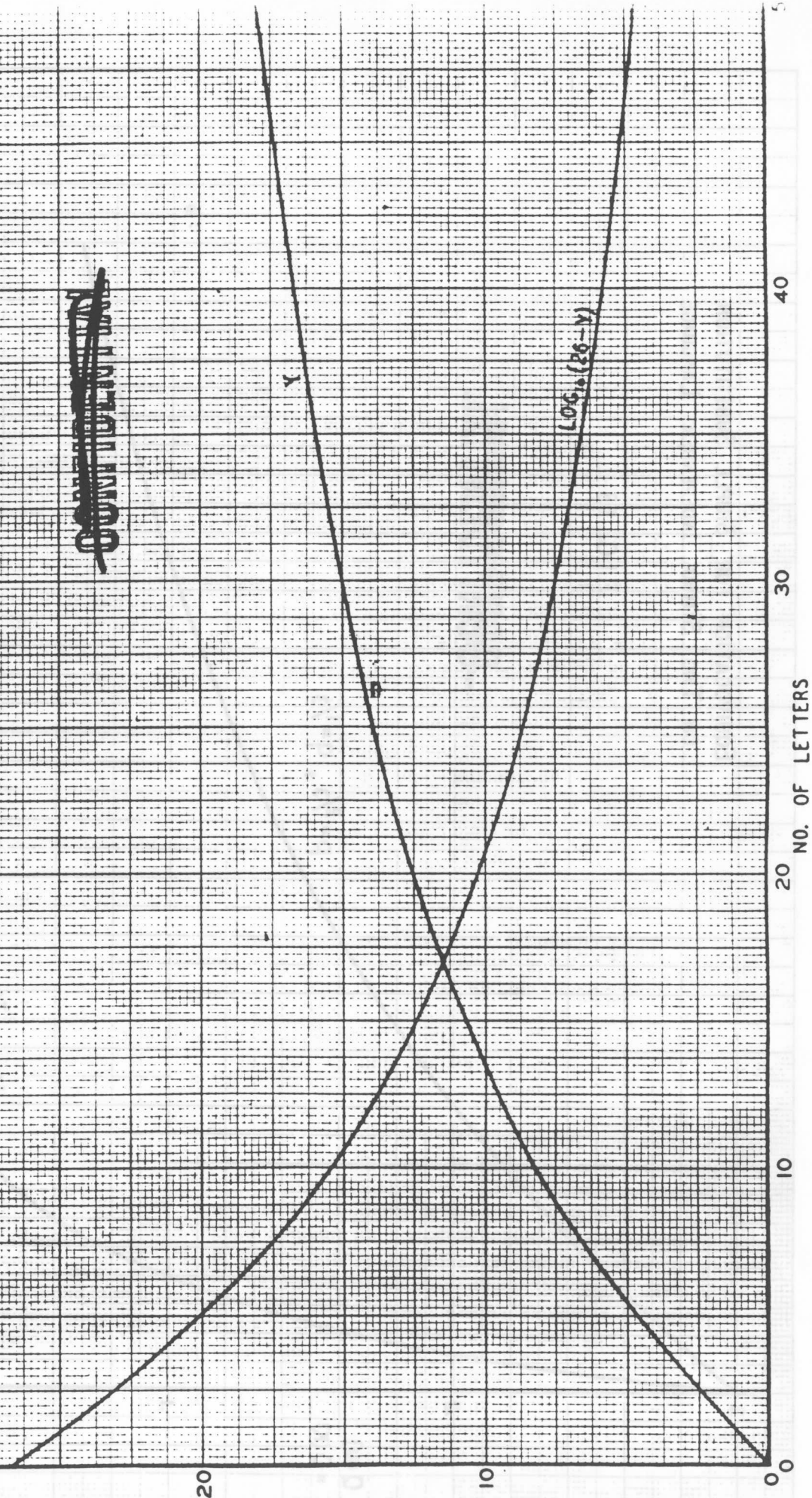
FIG. II

CONFIDENTIAL

Y-NO. OF DIFFERENT LETTERS

FIG. 12

~~CONFIDENTIAL~~



NO. OF LETTERS

EQUIVOCATION FOR SIMPLE SUBSTITUTION ON TWO SYMBOL INDEPENDENT LANGUAGE

FIG. 13

~~CONFIDENTIAL~~

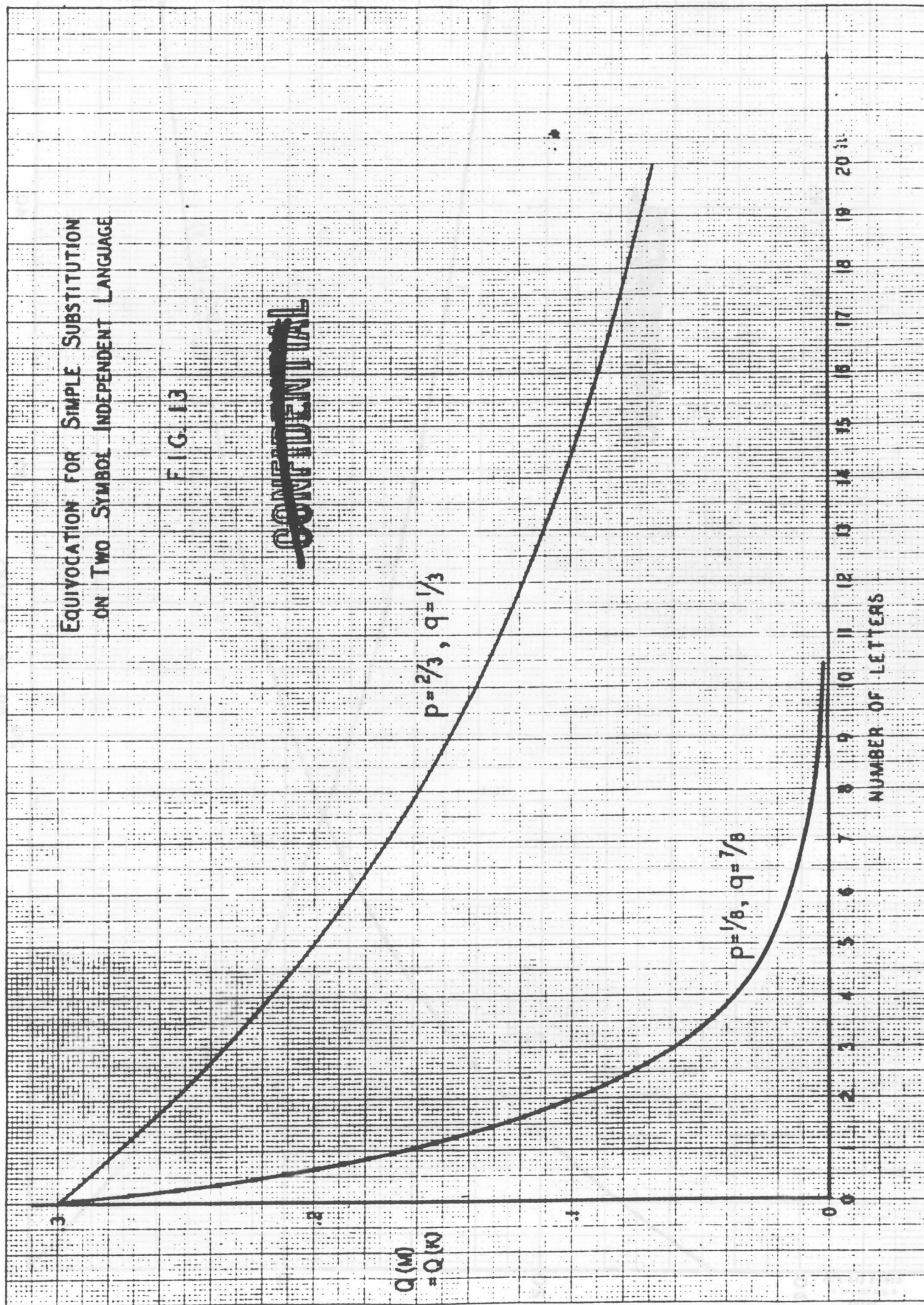
$Q(M)$
 $= Q(K)$

$p = \frac{2}{3}, q = \frac{1}{3}$

$p = \frac{1}{8}, q = \frac{7}{8}$

NUMBER OF LETTERS

20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 0



EQUIVOCATION FOR RANDOM CIPHER

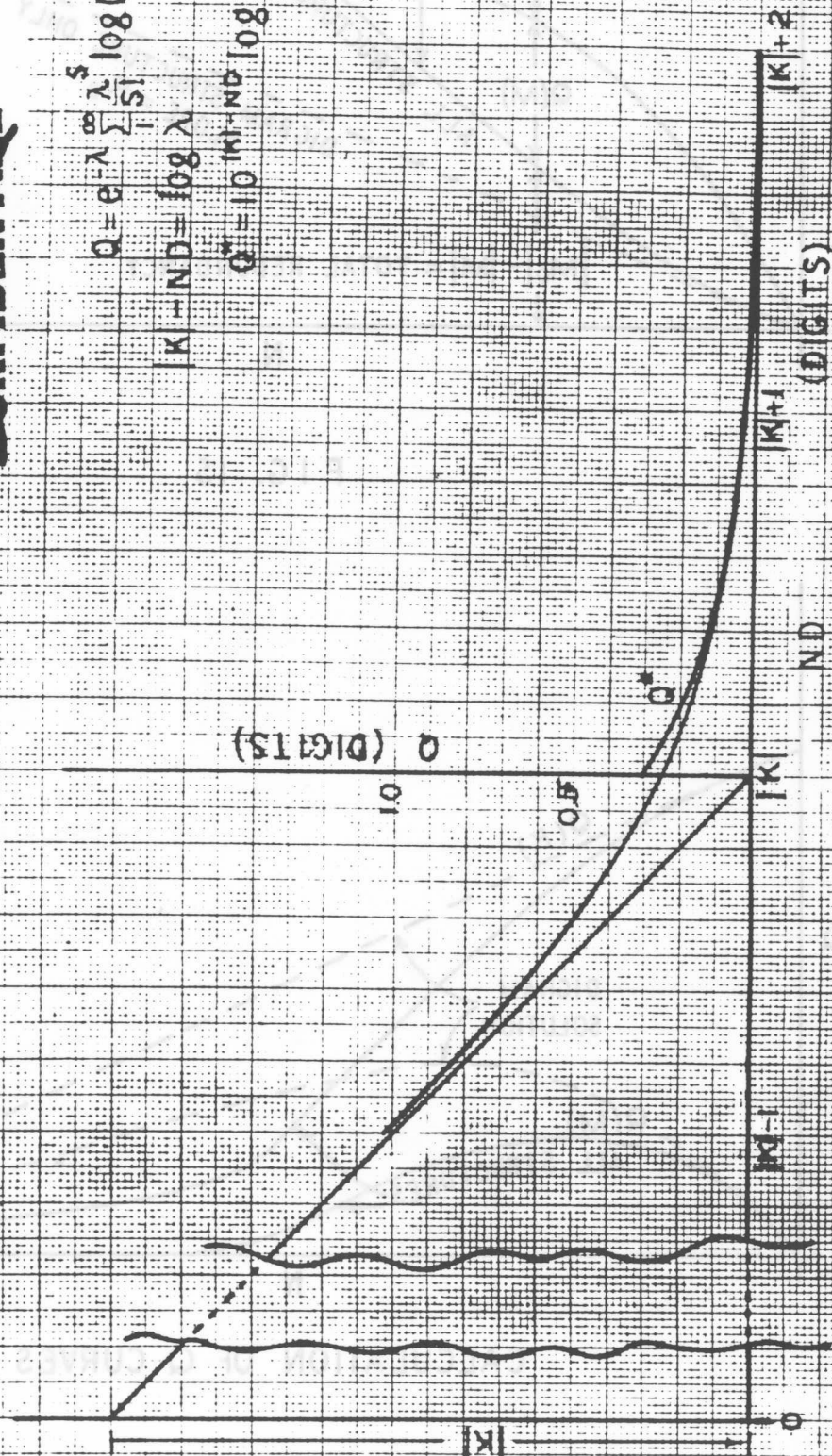
FIG. 14

~~CONFIDENTIAL~~

$$Q = e^{-\lambda} \sum_{s=1}^{\infty} \frac{\lambda^s}{s!} \log(s+1)$$

$$[K - ND] = \log \lambda$$

$$Q^* = 10^{K-ND} \log 2$$



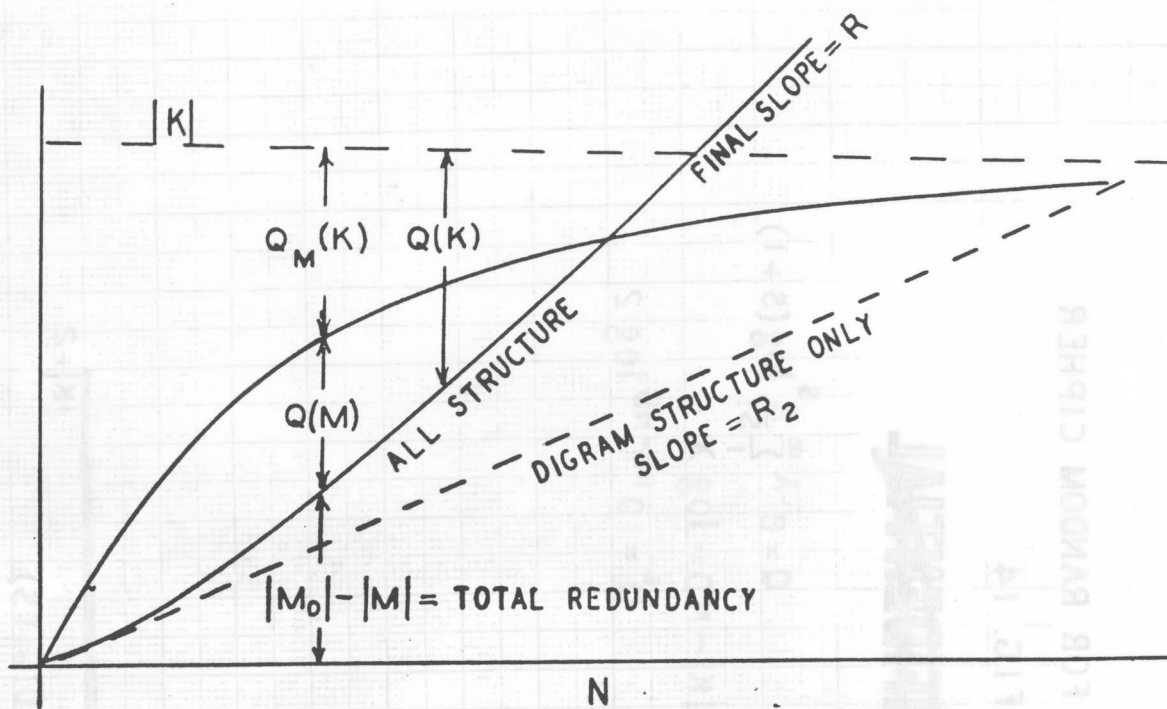
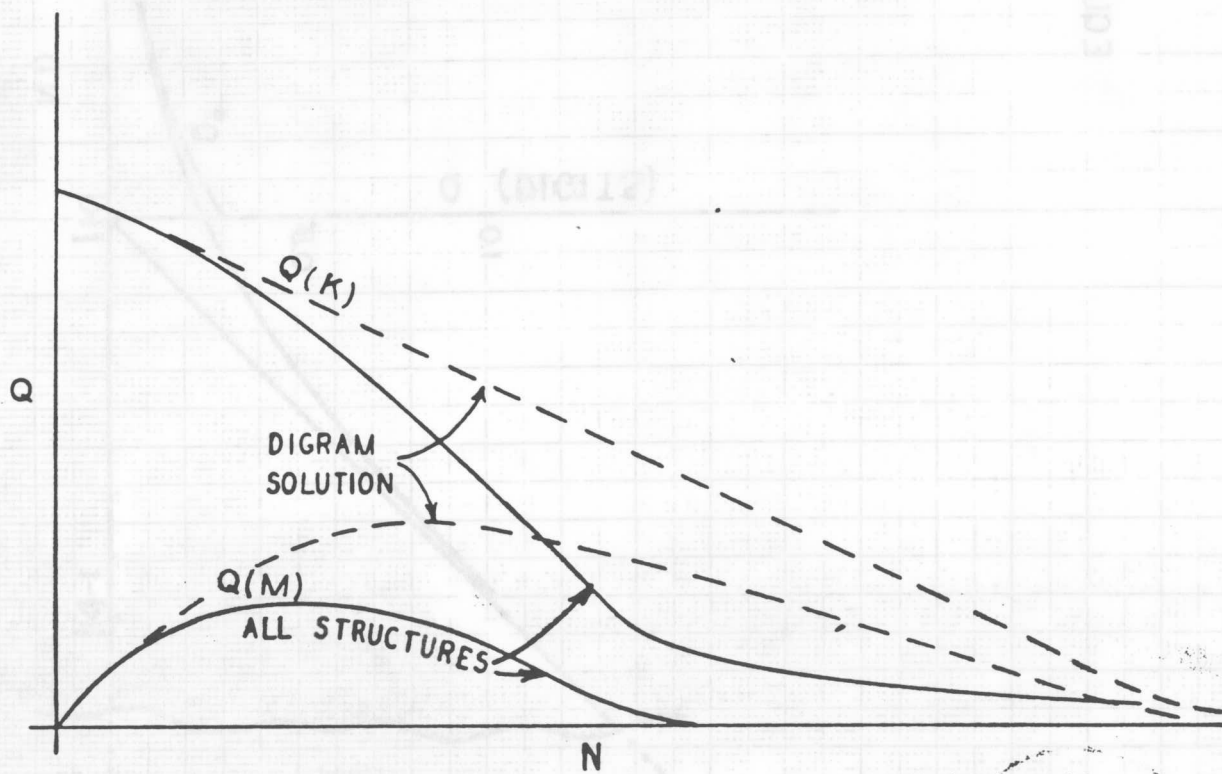


FIG. 15



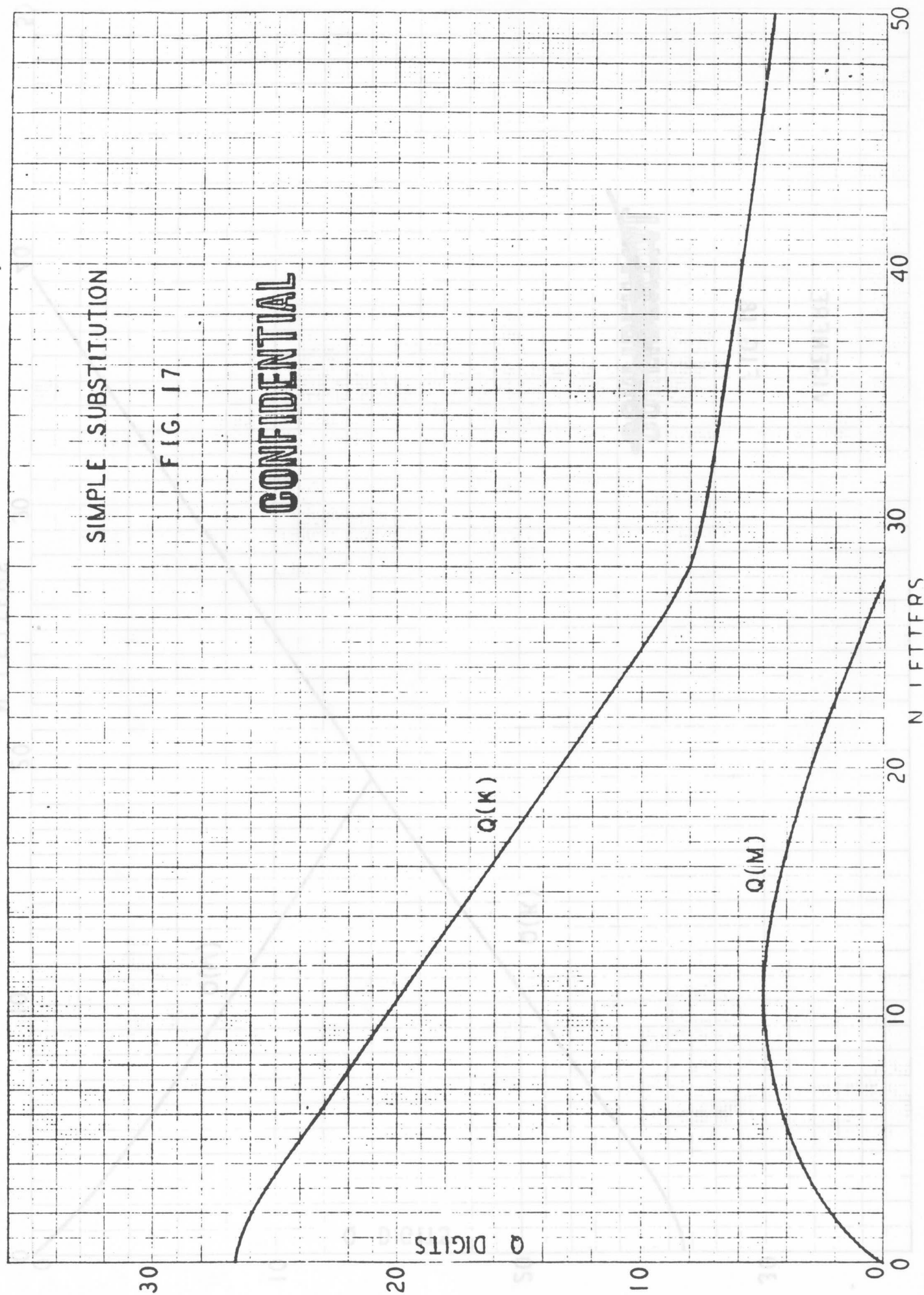
CALCULATION OF Q CURVES

FIG. 16

SIMPLE SUBSTITUTION

FIG. 17

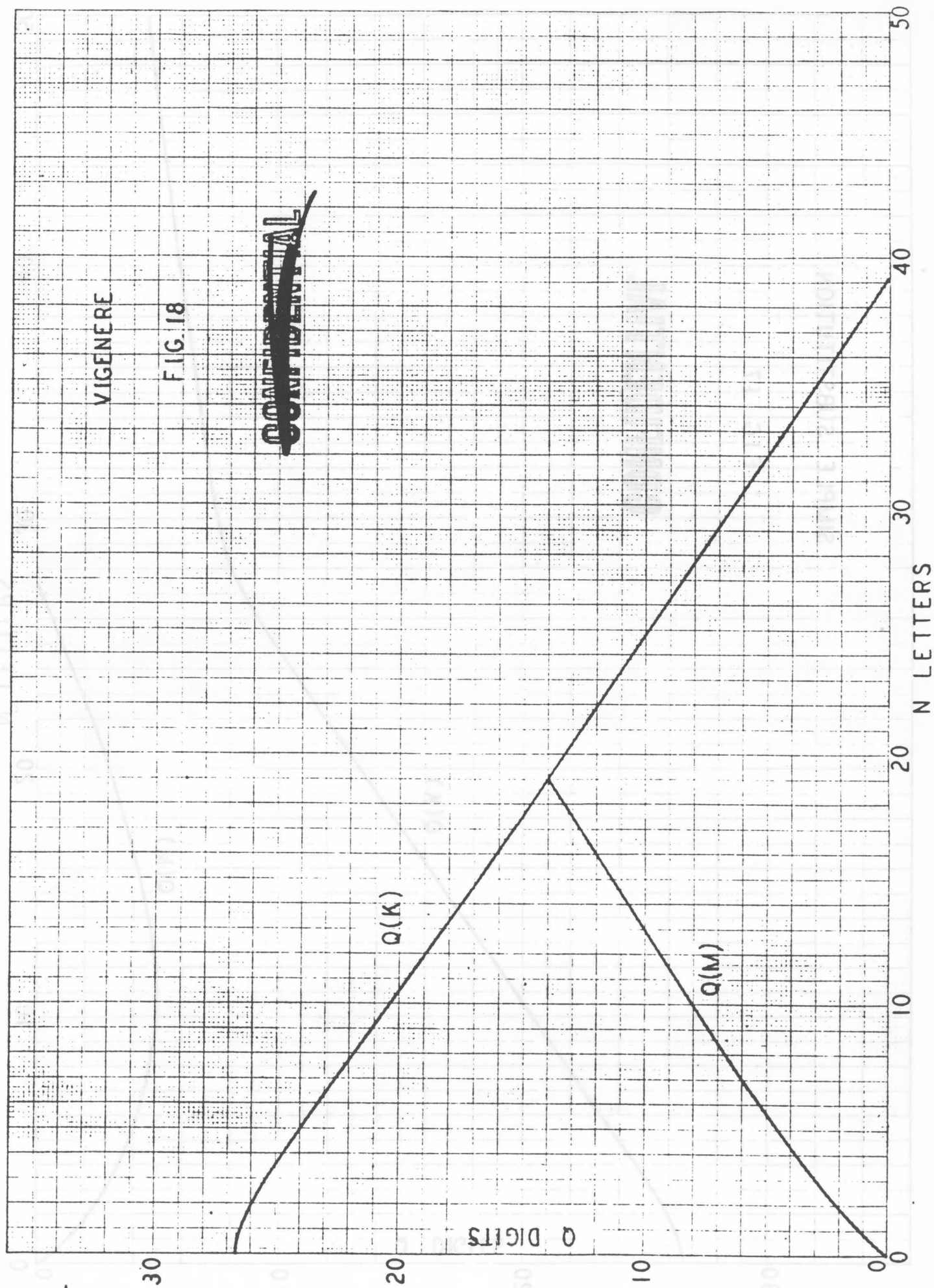
CONFIDENTIAL



VIGENERE

FIG. 18

~~CONFIDENTIAL~~



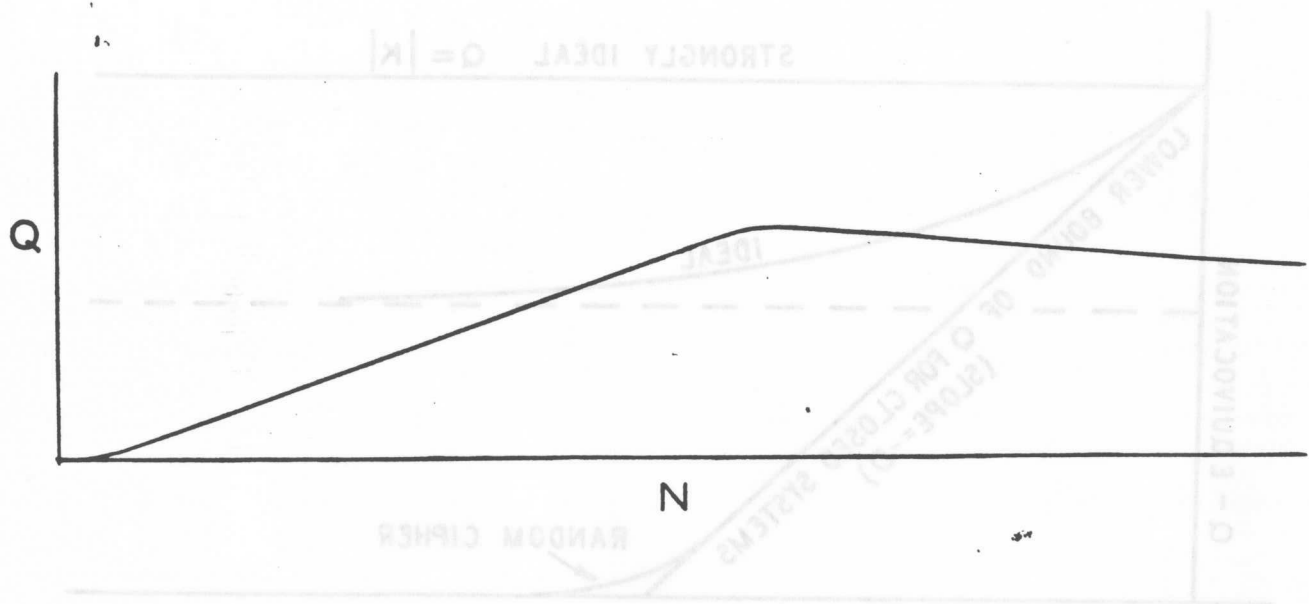
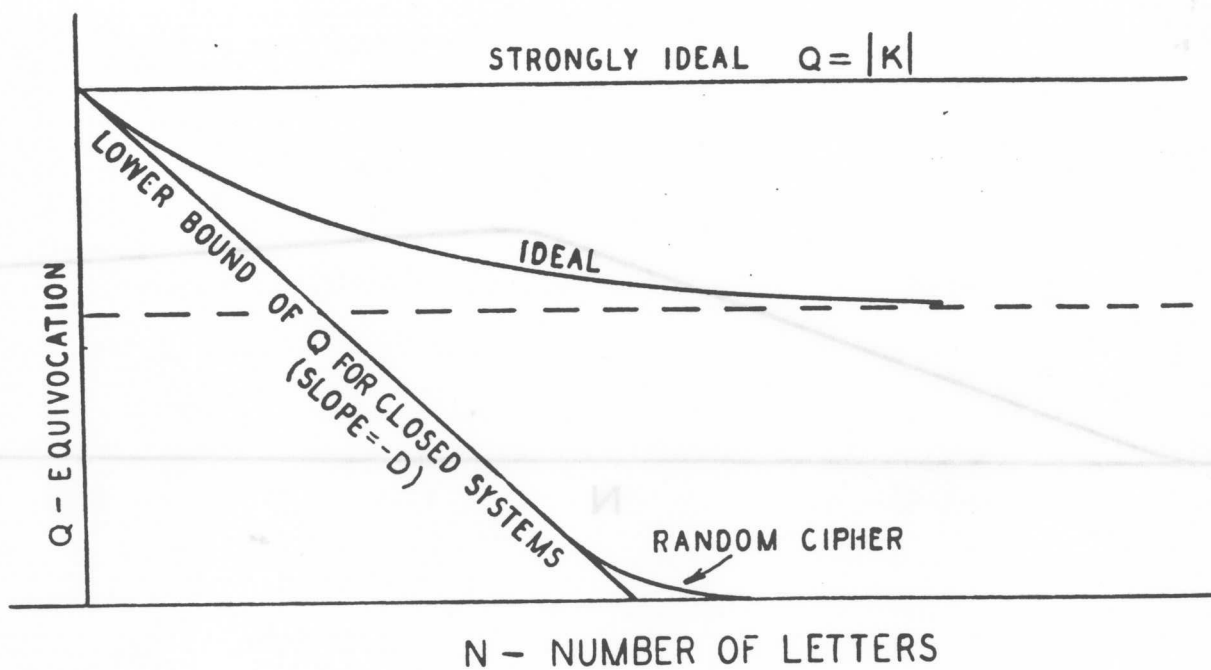


FIG. 19

CONFIDENTIAL



IDEAL CHARACTERISTICS

FIG. 20

~~CONFIDENTIAL~~

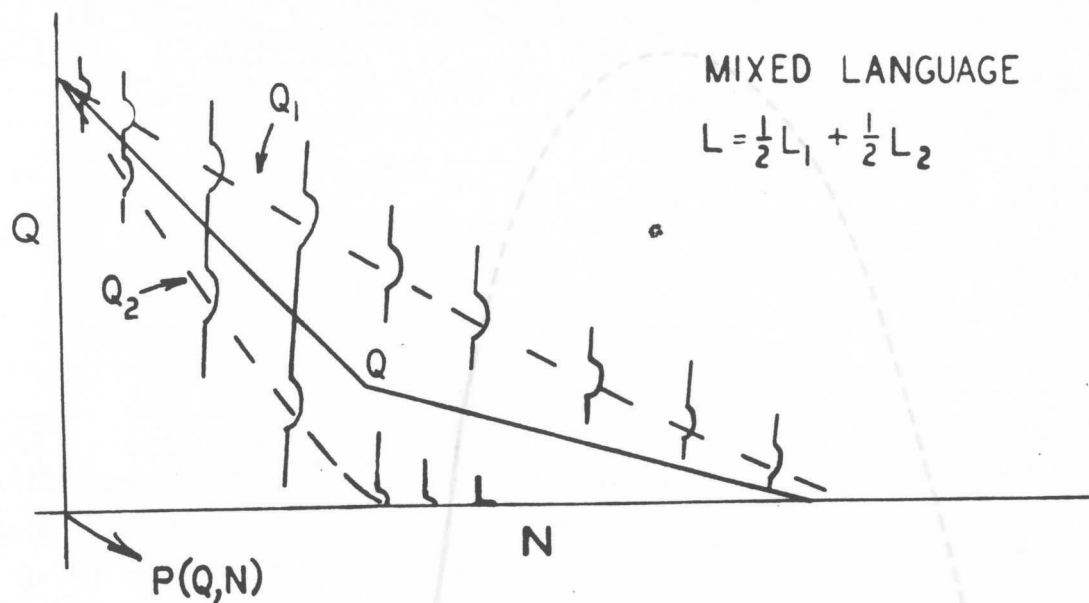


FIG. 21

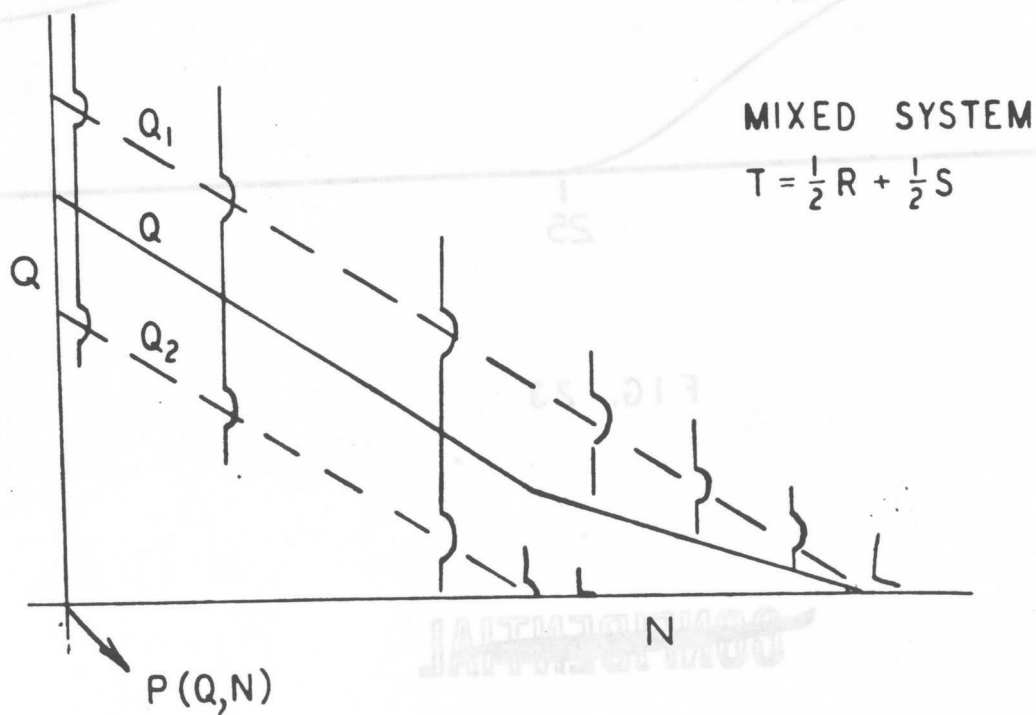


FIG. 22

~~CONFIDENTIAL~~

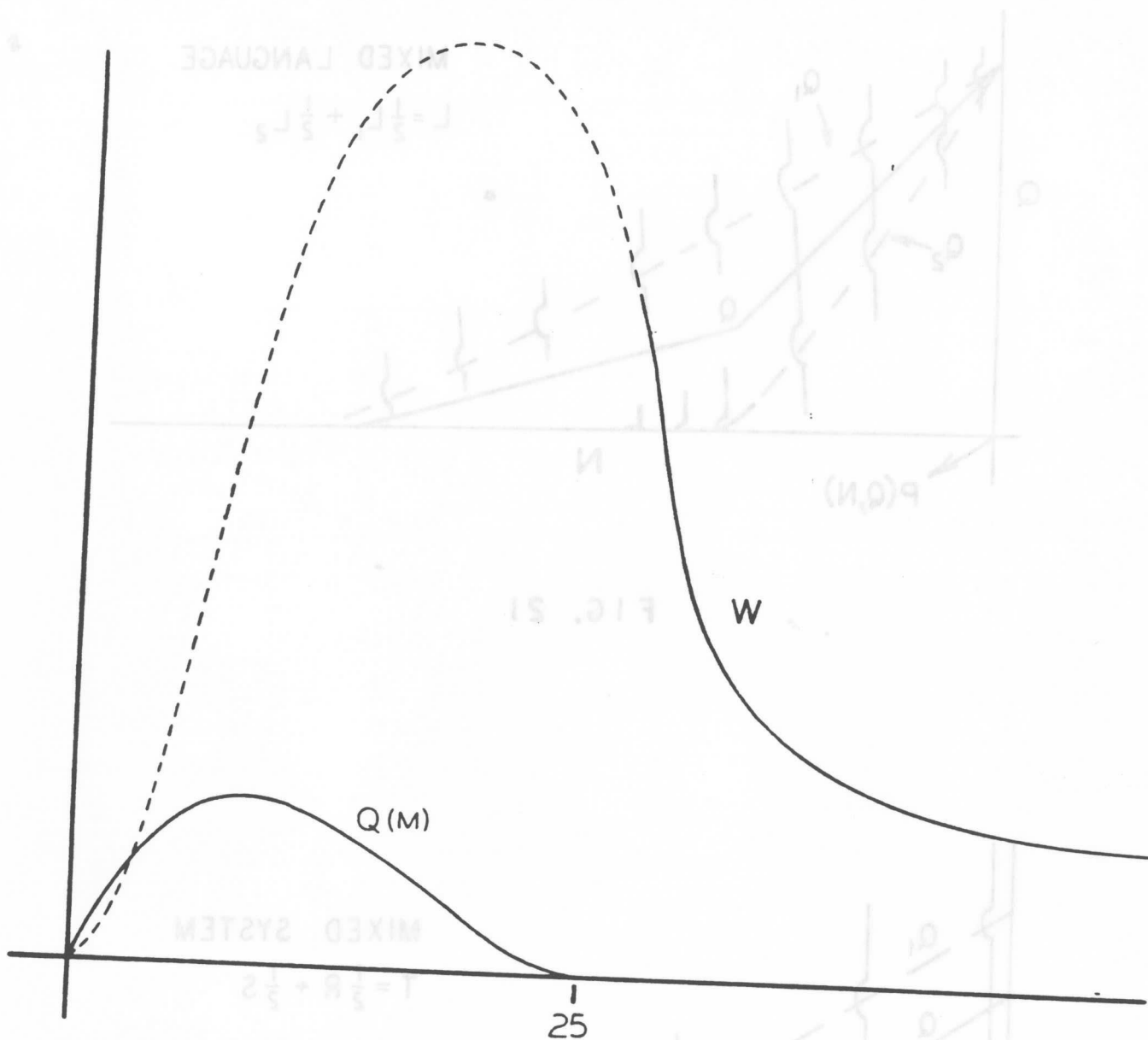
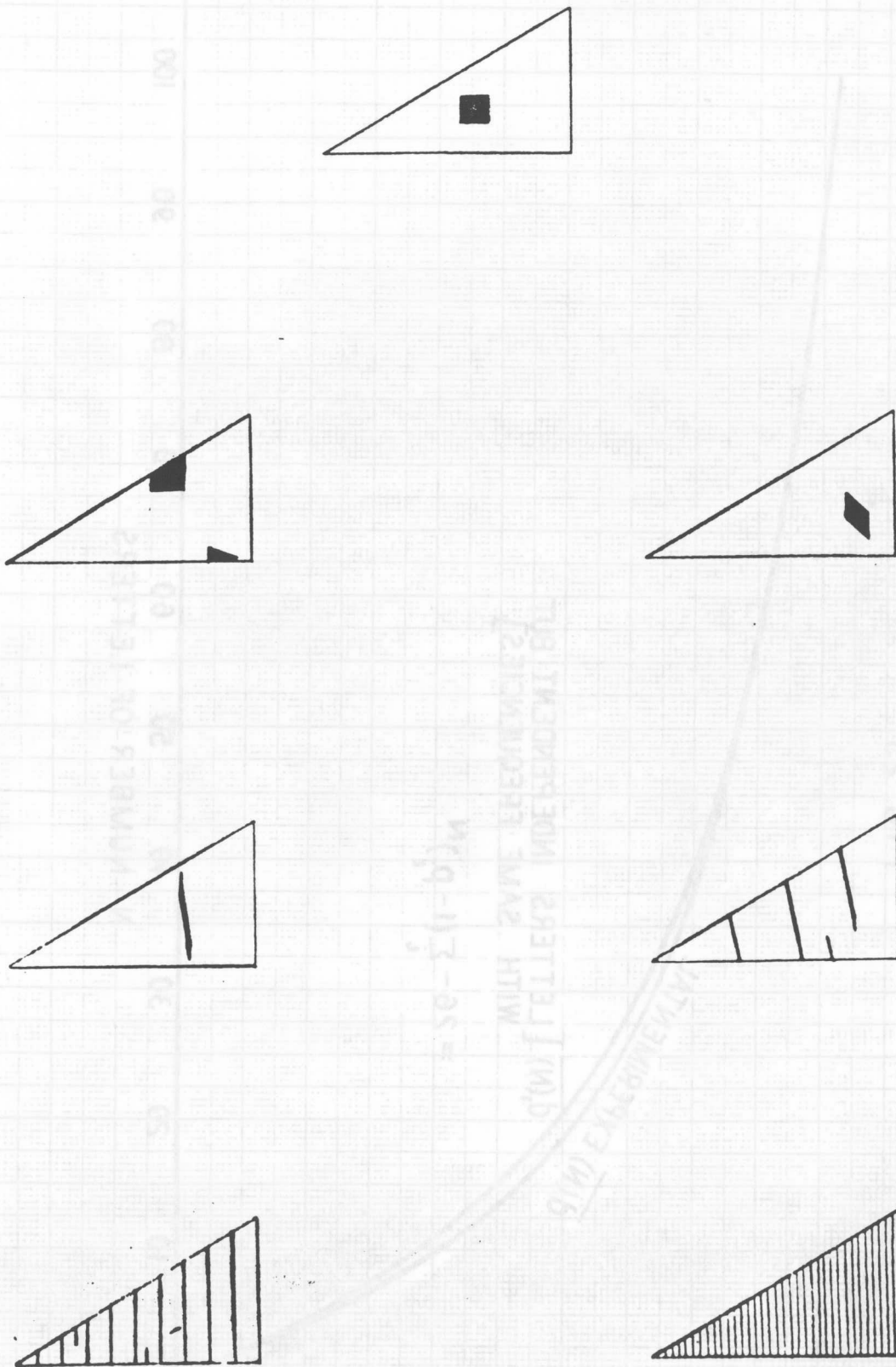


FIG. 23

~~CONFIDENTIAL~~



SIMPLE MIXING TRANSFORMATION.

FIG. 24

~~CONFIDENTIAL~~

EXPECTED NUMBER OF DIFFERENT
LETTERS IN N LETTERS OF TEXT

~~CONFIDENTIAL~~

FIG. 25

NUMBER OF DIFFERENT LETTERS

N NUMBER OF LETTERS

$d(N)$ EXPERIMENTAL

$d(N)$ { LETTERS INDEPENDENT BUT
WITH SAME FREQUENCIES }

$$= 26 - \sum_i (1 - p_i)^N$$

100

90

80

70

60

50

40

30

20

10

0

26

24

22

20

18

16

14

12

10

8

6

4

2

0

[26]

September 19, 1945-1125-CES-FG

Introduction.

In classical mechanics one considers situations where the state of a system is described by a set of numbers, the coordinates of the phase space of the system, and the dynamical behavior is controlled by a set of ordinary differential equations. Such a system is entirely determinate; the future is completely specified by the present state and the dynamical equations, since these differential equations have, in general, a unique solution passing through a given point.

In other branches of physics (heat flow, brownian motion, diffusion etc.) there are situations which can be called completely statistical. The path of a particle of gas is described only statistically and not determinate or mean behavior occurs. In this case one studies the flow of probability which is described by a partial differential equation of the heat flow type.

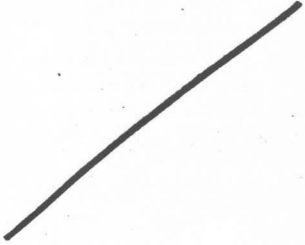
The present memorandum discusses a partial differential equation in which both effects occur--there is a definite "mean" motion of a system determinate in character, carrying its representative point through phase space in the classical manner with a superimposed statistical effect continually perturbing it from this path.

In such a case the future coordinates of the systems cannot be precisely predicted; only a probability distribution function can be determined for the future time whose value times the volume element dv is the probability that the system will be in the volume element dv around the point in question. For a short time the system is substantially determinate, the distribution being concentrated around a point which moves according to the determinate part of the equation. As the statistical effects come into play this distribution broadens out and in general approaches a limiting distribution which is independent of the initial state of the system.

In some respects the situation is similar to that in quantum mechanics, where systems are described only by probabilities (or more precisely by wave functions whose squared amplitudes are probabilities). There is this difference however; in quantum mechanics even the initial state cannot be precisely described due to the uncertainty principle. Conjugate variables cannot both be measured simultaneously with exactness. In the systems we consider here there are assumed to be no difficulties of this nature--all coordinates can be simultaneously and precisely measured. This corresponds to the difference in the fundamental equation from that of quantum mechanics--Schrödinger's equation is for the wave function ψ , while the equation considered here deals directly with the probability density. Thus the present work is adapted to "Molar" statistical situations.

This sort of analysis may be expected to apply to many problems where the actual situation is quite complicated but a partial theoretical analysis is possible. This partial analysis is used for the determinate part of the equation, and the other complex disturbing effects treated statistically. Such situations may occur in economics, sociology, history, etc. as well as in many engineering and physical problems.

G. R. Stibitz in a series of memoranda has considered a similar problem in connection with the stability of a periodically closed servo system. In his case the phase space of the system consisted of a set of discrete points, and the fundamental equation is a difference equation. In the case considered here (which was suggested by Stibitz' work) the variables are continuous and a differential equation is involved.



The Fundamental Equation.

In a determinate system with an n dimensional phase space, whose motion is described by differential equations, we have

$$\frac{dx^i}{dt} = r^i(x^1, x^2, \dots, x^n) \quad i = 1, 2, \dots, n \quad (1)$$

where the x^i are coordinates in the phase space and t is time.

If we start with a probability distribution of points in phase space

$$P(x^1, \dots, x^n, t)$$

giving the probability density in the differential volume element about x^1, \dots, x^n at time t , this distribution changes with time. Its motion is described by the partial differential equation

$$\frac{\partial P}{\partial t} + \sum_i \frac{\partial}{\partial x^i} (P r^i) = \nabla \cdot P \mathbf{r}$$

or in tensor notation

$$\frac{\partial P}{\partial t} + \frac{\partial}{\partial x^i} P r^i$$

This is evident if we think of P as a fluid density whose velocity field is \mathbf{r} .

Now suppose that as the representative point of the system moves about the phase space it is continually subject to small disturbances, which are of a probability type. Thus the system tends to follow the solution of (1) but is continually being disturbed by the probability effects, which may be thought of as something like molecular collisions of the surrounding gas

on a moving particle. We are interested in the limiting case where the disturbing effects are very rapid but very small. If we assume that the disturbance is homogeneous and isotropic, this can be represented by an additional term in the equation of the heat flow type

$$\kappa \nabla^2 P.$$

In the more general case certain directions may be preferred, and certain regions may have greater perturbation effects. Thus there will generally be a small ellipsoid of probability about each point, and a corresponding positive definite quadratic form

$$a^{ij}(x^1, \dots, x^n)$$

defined over the phase space. This form describes the local statistical perturbing effects, for each point.

The equation then assumes the form

$$\frac{\partial P}{\partial t} - \frac{\partial}{\partial x^i} a^{ij} \frac{\partial P}{\partial x^j} + \frac{\partial}{\partial x^j} P x^j \quad (1)$$

where repeated indices are summed.

This partial differential equation governs the flow of probability in the phase space. With an ensemble of systems distributed at $t = 0$ according to $P_0(x^i)$ the distribution at a later time t_1 is the solution of (1) for $t = t_1$.

The equation (1) is linear and of parabolic type (in t). In the x^i it is elliptical, since a^{ij} is positive definite.

Conservation and Limiting Properties.

The total probability in all phase space remain constant, for if we let

$$U = \int P \, dv$$

$$\frac{dU}{dt} = \int \frac{\partial P}{\partial t} \, dv = \int \frac{\partial}{\partial x^i} \left(a^{ij} \frac{\partial P}{\partial x^j} + P r^i \right) \, dv$$

$$= \int \left(a^{ij} \frac{\partial P}{\partial x^j} + P r^i \right) N_i \, ds = 0$$

the integral being over a sufficiently large surface, and N_i the unit normal.

If a^{ij} is positive definite and both a^{ij} and r^i are continuous in the phase space the distribution P approaches a unique limit as $t \rightarrow \infty$. This limit is either zero everywhere, the probability retreating to infinity or a definite limiting distribution P^* with

$$\int P^* \, dv = \int P \, dv$$

for any t .

The limiting distribution must satisfy the elliptical equation obtained by setting $\frac{\partial P}{\partial t} = 0$,

$$\frac{\partial}{\partial x^i} a^{ij} \frac{\partial P}{\partial x^j} + \frac{\partial}{\partial x^i} P r^i = 0$$

To show that the distribution approaches a limit let P_1 and P_2 be two different solutions of (1). Then the difference $Q = P_1 - P_2$ also satisfies the equation and Q is positive in one region R and negative in the remainder of the space. Consider the quantity

$$U = \int_R Q \, dv$$

U must decrease for

$$\dot{U} = \frac{d}{dt} \int_R Q \, dv = \int_R \dot{Q} \, dv + \int_S Q \cdot V \, d\sigma$$

where S is the surface of the region R and V is the outward velocity of this surface. Since Q vanishes on the surface, the second term is zero, and the first is

$$\dot{U} = \int_R \left[\frac{\partial}{\partial x^1} a^{1j} \frac{\partial Q}{\partial x^j} + \frac{\partial}{\partial x^1} r^1 Q \right] dv$$

These are volume integrals of divergences and transform by the usual theorems into surface integrals

$$\dot{U} = \int_S \left[a^{1j} \frac{\partial Q}{\partial x^j} N_1 + r^1 Q N_1 \right] d\sigma$$

the second term again vanishes since $Q = 0$ on S . In the first term N_1 is in the direction of $\frac{\partial Q}{\partial x^1}$ so at any point we have

$$K a_{1j} \frac{\partial q}{\partial x^j} \frac{\partial q}{\partial x^i} \leq 0$$

so $\dot{U} \leq 0$ and the discrepancy between P_1 and P_2 is decreasing.
Thus any initial distribution approaches the same limit.

Discontinuities in r^1 .

If a^{1j} is continuous, but r^1 has a discontinuity, P will be continuous, and the vector $\frac{\partial P}{\partial x^1}$ discontinuous.

The amount of this discontinuity is given by

$$a^{1j} (r_j - \bar{r}_j) = - (r^1 - \bar{r}^1) P$$

where the barred and unbarred letters refer to the two sides of the discontinuity. Thus

$$\frac{r_1 - \bar{r}_1}{P} = - (r^1 - \bar{r}^1) a_{1j}$$

This relation allows one to fit solutions of the steady state equation on the two sides of such discontinuities.

In the simplest one dimensional case we have

$$\frac{\frac{dr}{dx} - \frac{\bar{r}}{dx}}{P} = - \frac{(r - \bar{r})}{a}$$

Immediate Behavior of a Localized System.

If we start with a "spike" of probability localized at one point, the immediate behavior can be described in simple terms. Near this point we may assume a^{ij} and r^i to be constant. Due to the r^i the spike starts moving with a velocity r^i , while the probability term a^{ij} spreads it out. If we ^{change} carry variables from x^i to

$$y^i = x^i + r^i t$$

we have

$$\frac{\partial P}{\partial x^i} = \frac{\partial \bar{P}}{\partial y^i} \quad , \quad \frac{\partial P}{\partial t} = \frac{\partial \bar{P}}{\partial y^i} r^i + \frac{\partial \bar{P}}{\partial t}$$

and the equation becomes

$$\frac{\partial P}{\partial t} + \frac{\partial \bar{P}}{\partial y^i} r^i = a^{ij} \frac{\partial^2 \bar{P}}{\partial y^i \partial y^j} + r^i \frac{\partial \bar{P}}{\partial y^i}$$

or

$$\frac{\partial \bar{P}}{\partial t} = a^{ij} \frac{\partial^2 \bar{P}}{\partial y^i \partial y^j}$$

This is the equation for heat flow in an anisotropic medium.

Thus in the y^i coordinate the spike diffuses out into a gaussian distribution with quadratic form a^{ij} , for the first short interval of time

$$P = \frac{|A_{ij}|^{\frac{1}{2}}}{\sqrt{\pi} (2\pi)^{\frac{n}{2}}} \exp \left(-\frac{1}{2t} A_{ij} y^i y^j \right)$$

where A_{ij} is the inverse form of a^{ij} .

Linear Velocity Field and Homogeneous Statistical Effects.

One particular case of interest is that in which the velocity field is linear and the statistical effects homogeneous in the space. For a one dimensional phase space, the equation (1) then assumes the form

$$a \frac{\partial^2 P}{\partial x^2} + b \frac{\partial}{\partial x} (xP) = \frac{\partial P}{\partial t} \quad (2)$$

A general solution for this case has been found. It may be described as follows. If the initial distribution is a δ function, so the system (or ensemble) is known to have a definite value of x at $t = 0$, say β , then at t_1 the distribution is normal. The center of this normal distribution is at

$$\bar{x} = \beta e^{-bt}$$

and its standard deviation σ is given by

$$\sigma^2 = \frac{a}{b} (1 - e^{-2bt})$$

Thus the mean decreases along the same curve as the system would follow were the statistical effects absent. The variance σ^2 increases exponentially to a limiting value a/b with half the time constant.

To prove that this is the solution it is only necessary to substitute in the equation (2). As $t \rightarrow \infty$ the distribution approaches a normal one centered on zero with $\sigma^2 = a/b$.

Handwritten signature

71

In symbols the solution is

$$P_1(x,t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x-\bar{x}}{\sigma} \right)^2}$$

$$\bar{x} = \beta(1 - e^{-bt})$$

$$\sigma^2 = \frac{\sigma}{b} (1 - e^{-2bt})$$

With an arbitrary initial distribution $P_0(x)$ the solution can be written as an integral using the method of superposition of heat flow problems.

$$P(x,t) = \int_{-\infty}^{\infty} P_0(\beta) P_1(x,t) d\beta$$

The same general results hold in the N dimensional case when r^i is a linear form and a^{ij} is constant. A "spike" of probability broadens into a normal distribution, the center following the determinate trajectory and the quadratic form which takes the place of the standard deviation increases exponentially toward a definite limit. The evaluation of the constants is much more complicated in this case however. The equations for the final distribution are given in the appendix.

Autocorrelation for Linear System in Statistical Equilibrium.

If in the one dimensional linear case we start with a normal distribution centered on zero with $\sigma^2 = \frac{a}{b}$, the distribution remains stably with the form. An individual system executes statistical motion about zero and the ensemble of systems produces an ensemble of time series. This ensemble can be seen to be equivalent to thermal noise which has been passed through a filter with transfer characteristic

$$Y = \frac{1}{b + j\omega}$$

leading to a power spectrum for the noise

$$P(\omega) = A \frac{1}{b^2 + \omega^2}$$

To show this, the autocorrelation may be calculated. Systems whose value at $t = 0$ is β have a normal distribution centered about $\beta e^{-b|t_1|}$ at t_1 . Hence for these, the mean value of $f(t) f(t + t_1)$ is

$$\beta^2 e^{-b|t_1|}$$

The distribution at $t = 0$ is normal with $\sigma^2 = \frac{a}{b}$.

Hence

$$\int P(\beta) \beta^2 e^{-b|t_1|} d\beta = \frac{b}{a} e^{-b|t_1|}$$

and this is the autocorrelation.

The power spectrum is the Fourier transform of this autocorrelation

$$P(\omega) = \frac{b}{a(1 + \frac{\omega^2}{b^2})} = \frac{b^3}{a(\omega^2 + b^2)}$$

Final Distribution in the General one dimensional case.

In the general one dimensional case we have

$$\frac{\partial P}{\partial t} = \frac{\partial}{\partial x} \left(a(x) \frac{\partial P}{\partial x} + f(x) P \right)$$

$a(x) \geq 0$. In the steady state

$$\frac{d}{dx} \left[a(x) \frac{dP}{dx} + f(x) P \right] = 0$$

$$a(x) \frac{dP}{dx} + f(x) P = 0$$

assuming $P, \frac{dP}{dx} \rightarrow 0$ as $x \rightarrow \pm \infty$, $a = 0$

and $a(x) \frac{dP}{dx} + f(x) P = 0$

$$\frac{dP}{P} = - \frac{f}{a} dx$$

$$\ln P = - \int \frac{f}{a} dx + K$$

$$P = A e^{- \int \frac{f}{a} dx}$$

$$\frac{f}{a} = \varphi(x)$$

$$= A e^{- \int_0^x \varphi(x) dx}$$

where A is determined by the condition $\int P = 1$.

For stability, $P \rightarrow 0$ as $x \rightarrow \pm \infty$

therefore it is necessary that $\int_0^x \varphi(x) dx \rightarrow \pm \infty$ as $x \rightarrow \pm \infty$

In the special case where

$$\varphi(x) = -\alpha \quad x > 0$$

$$\varphi(x) = +\alpha \quad x < 0$$

we obtain as the final stationary solution

$$P = \frac{1}{2a} e^{-a|x|}$$

a pair of exponentials decreasing toward $\pm \infty$.

C. E. SHANNON

Appendix. n dimensional linear case.

If

$$\frac{dx^j}{dt} = \beta_j^i x^i$$

and a^{ij} is constant

$$\begin{aligned} \frac{\partial P}{\partial t} &= a^{ij} \frac{\partial^2 P}{\partial x^i \partial x^j} + \frac{\partial}{\partial x^j} \beta_j^i x^i P \\ &= a^{ij} P_{,ij} + \beta_j^i x^i P_{,j} + P \beta_j^j \end{aligned}$$

assume $P = \exp - \frac{1}{2} c_{ij} x^i x^j$

$$P_{,i} = -P c_{is} x^s$$

$$P_{,ij} = P (c_{is} x^s c_{jr} x^r - c_{ij})$$

To satisfy $\frac{\partial P}{\partial t} = 0$ we must then have

$$a^{ij} (c_{is} x^s c_{jr} x^r - c_{ij}) - \beta_j^i x^i c_{is} x^s + \beta_j^j = 0$$

This requires that

$$a^{ij} c_{ij} = \beta_j^j$$

and also $2a^{ij} c_{is} c_{jr} = \beta_j^i c_{is} + \beta_s^i c_{ir}$

Hence

$$2a^{is} c_{is} = \beta_s^i c_{is} c^{rs} + \beta_s^i c_{is}$$

$$2a^{is} = \beta_s^i c^{rs} + \beta_s^i c^{rs}$$

$$= c^{rs} [\beta_s^i c_{rk} + \beta_r^i c_{sk}]$$

(2)

In coordinate which diagonalize β_j^1 we have from (2)

$$x_2^{nt} = \lambda_2 \bar{c}^{nt} + \lambda_3 \bar{u}^{nt}$$

thus

$$\bar{c}^{nt} = \frac{x_2^{nt}}{\lambda_2 + \lambda_3}$$

DATA SMOOTHING AND PREDICTION IN FIRE-CONTROL SYSTEMS

By R. B. Blackman, H. W. Bode, and
C. E. Shannon *

THE PROBLEM of data smoothing in fire control arises because observations of target positions are never completely accurate. If the target is located by radar, for example, we may expect errors in range running from perhaps 10 to 50 yards in typical cases. Angular errors may vary from perhaps one to several mils, corresponding at representative ranges, to yardage errors about equal to those mentioned for range. Similar figures might be cited for the errors involved in optical tracking by various devices. Evidently these errors in observation will generate corresponding errors in the final aiming orders delivered by the fire-control system.

A data-smoothing device is a means for minimizing the consequences of observational errors by, in effect, averaging the results of observations taken over a period of time. The simplest example of data smoothing is furnished by artillery fire at a fixed land target. Here the principal parameter is the range to the target. While individual determinations of the range may be somewhat in error, a reliable estimate can ordinarily be obtained by taking the simple average of a number of such observations. This example, however, is scarcely a representative one for problems in data smoothing generally. The errors involved are small and the averaging process is an elementary one. Moreover, the data-smoothing process is not of very decisive importance in any case, since any errors which may exist in the estimated range can normally be wiped out merely by observing the results of a few trial shots.

More representative problems in data smoothing arise when we deal with a moving target. In this case errors in observational data may be much more serious, since they determine not only the present position of the target but also the rates used in calculating how much the target will move during the time it takes the projectile to reach it. An illustration is furnished by anti-aircraft fire against

distant airplanes. Suppose, for example, that in observing the target's position we make two errors of opposite sign and a second apart, of 25 yards each. Then the apparent motion of the airplane is in error by 50 yards per second. Since the time of flight of an anti-aircraft shell in reaching its target may be as high as 80 seconds or more, such an error might produce a miss of the order of 1 mile. It is clear that in any comparable situation the effect of observational errors in determining the target rate will be much greater than the position error alone would suggest, and the function of the data-smoothing network in averaging the data so that even moderately reliable rates can be obtained as a basis for prediction becomes a critically important one.

Aside from magnifying the consequences of small errors in target position, the motion of the target complicates the data-smoothing problem in two other respects. The first is the fact that it gives us only a brief time in which to obtain suitable firing orders. The total engagement is likely to last for only a brief time, and in any case it is necessary to make use of the data before the target has time to do something different. Thus the averaging process cannot take too long. The second complication results from the fact that the true target position is an unknown function of time rather than a mere constant. Thus many more possibilities are open than would be the case with fixed targets, and the problem of averaging to remove the effects of small errors is correspondingly more elusive.

The intimate relation between data smoothing and target mobility explains why the data-smoothing problem is relatively new in warfare. The problem emerged as a serious one only recently, with the introduction of new and highly mobile military devices. The airplane is, of course, the archetype of such mobile instruments, and we have already mentioned the data-smoothing problem as it appears in anti-aircraft fire. Since the relative velocity of airplane and ground is the same whether we station ourselves on one or the other, however, the

* Bell Telephone Laboratories.

mobility of the airplane produces essentially the same sort of problem in the design of bombsights also. Another field exists in plane-to-plane gunnery. Although they are somewhat slower, the mobility of such vehicles as tanks and torpedo boats is still considerable enough to create a serious problem here also. Future examples may be centered largely on robot missiles. It is interesting to notice that a guided missile may present a problem in data smoothing either because it belongs to the enemy, and is therefore something to shoot at, or because it belongs to us, and requires smoothing to correct errors in the data which it uses for guidance. The tendency to higher and higher speeds in all these devices must evidently mean that fire control generally, and data smoothing as one aspect of fire control, must become more and more important, unless war making can be ended.

Very mobile instruments of war, such as the airplane, began to make their appearance in World War I, but there was insufficient time during that war to make much progress with the fire-control problems which such instrumentalities imply. In the interval between World War I and World War II, however, a considerable number of fire-control devices, such as bombsights and antiaircraft computers, were developed. The principal attention in the design of these devices, however, was on the kinematical aspects of the situation. Although a number of them included fairly successful methods of minimizing the effects of observational errors,^b it seems fair to say that in the interval between the two wars there was no general appreciation of the existence of the data-smoothing problem as such.

It follows that the theory of data smoothing advanced in this monograph is the result principally of experience gained in World War II. More specifically, it is the product of the ex-

perience of the authors with a series of projects, largely sponsored by Division 7 of NDRC, concerned with the design of electrical antiaircraft directors. In addition, it draws largely on the results of a number of other investigations, also NDRC sponsored. The possible key importance of data smoothing in the design of fire-control systems was recognized by Division 7 early in the course of its activities and the emphasis placed upon it in a number of projects led to the accumulation of a much larger body of results than might otherwise have been obtained.

Data smoothing is developed here in terms of concepts familiar in communication engineering. This is a natural approach since data smoothing is evidently a special case of the transmission, manipulation, and utilization of intelligence. The other principal, and perhaps still more fundamental, approach to data smoothing is to regard it as a problem in statistics. This is the line followed in the classic work¹ by Norbert Wiener.^c For reasons which are brought out later, Wiener's theory is not used in the present monograph as a basis for the actual design of data-smoothing networks. Because of its fundamental interest, however, a sketch of Wiener's theory is included. The authors' apologies are due for any mutilation to the theory caused by the attempt to simplify it and compress it into a brief space.

The present monograph falls roughly into two dissimilar halves. The first half, consisting of the first three or four chapters, includes a discussion of the general theoretical foundations of the data-smoothing problem, the best established ways of approaching the problem, the assumptions they involve, and the authors' judgment concerning the assumptions which best fit the tactical facts. In this part may also be included the last chapter, which contains a fragmentary discussion of alternative data-smoothing possibilities lying outside the main theoretical framework of the monograph.

The rest of the monograph is concerned with the technique of designing specific data-smoothing structures. A fairly elaborate and detailed treatment is given here, in the belief that the

^b Most of these solutions depended upon the use of special types of tracking systems. Examples are found in the use of regenerative tracking in bombsights and antiaircraft computers or in the determination of rates from a precessing gyroscope or an aided laying mechanism in an antiaircraft tracking head. So far as their effect on the data-smoothing characteristics of the overall circuit is concerned, these devices are equivalent to simple types of smoothing networks inserted directly in the prediction system. This is discussed in more detail under the heading "Exponential Smoothing," Section 10.1.

^c Wiener is also responsible for providing tools which permit the gap between the statistical and communication points of view to be bridged.

problem of actually realizing a suitable data-smoothing device is, in some ways at least, as difficult as that of deciding what the general properties of such a device should be. The technique, as given, draws heavily upon the highly developed resources of electric network theory. For this reason the discussion is couched entirely in electrical language, although the authors realize, of course, that equivalent nonelectrical solutions may exist. For the benefit of readers who may not be familiar with network theory, the monograph includes an appendix summarizing the principles most needed in the main text.

Two further remarks may be helpful in understanding the monograph. The first concerns the relation between data smoothing and the overall problem of prediction in a fire-control circuit. These two are coupled together in the title of the monograph, and it is clear that the connection between them must be very close, since, as we saw earlier, small irregularities in input data are likely to be serious only as they affect the extrapolation used to determine the future position of a moving target. In the statistical approach, in fact, data smoothing and prediction are treated as a single problem and a single device performs both operations.

In the attack which is treated at greatest length in the monograph a certain distinction between data smoothing and prediction can be made. To simplify the exposition as much as possible, the explicit discussion in the monograph is directed principally at data smoothing. This, however, is not intended to suggest that there is any real cleavage between the two problems or that the analysis as developed in the monograph does not also bear, by implication, upon the prediction problem. Any theory of data smoothing must rest ultimately upon some hypothesis concerning the path of the target, and the exact statement of the assumptions to be made is in many ways the most important as well as the most difficult part of the problem. The same assumptions, however, are also involved in the extrapolation to the future position of the target. It is thus impossible to solve the data-smoothing problem without also implying what the general nature of the prediction process will be. For example, the formulation given in Chapter 9 amounts to

the assumption that the target path is specified by a set of geometrical parameters corresponding to components of velocity, acceleration, etc. The data-smoothing process centers about the problem of obtaining reliable values for these parameters. To obtain a complete prediction thereafter, it is merely necessary to multiply the parameter values thus obtained by suitable functions of time of flight and add the results to the present position of the target.

The other general remark concerns the tactical criteria used in evaluating the performance of a data-smoothing system. This turns out to be one of the most important aspects of the whole field. It is assumed here that the tactical situation is similar to that of anti-aircraft fire against high-altitude bombers in World War II. The defense can be regarded as successful if only a fairly small fraction of the targets engaged are destroyed. On the other hand, the lethal radius of the anti-aircraft shell is so small that it is also quite difficult to score a kill. Under these circumstances we are interested only in increasing the number of very well aimed shots.

When we combine these assumptions with the path assumptions described in Chapter 9 we are led to the data-smoothing solution formulated here, in preference to the solution obtained with the statistical approach. On the other hand, we might equally well envisage a situation in which the target contained an atomic bomb or some other very destructive agent, so that it becomes very important to intercept it, while the lethal radius of the anti-aircraft missile is correspondingly increased, so that great accuracy is not needed for a kill. In this situation our interest would be focused on the problem of minimizing the probability of making large misses, and the solution furnished by the statistical approach would be approximately the best obtainable.⁴

⁴ In fairness to the statistical solution it should be pointed out that it is also the best obtainable, without regard to the lethal radius of the shell, if we replace the path assumptions made in Chapter 9 by a "random phase" assumption. The path assumptions in Chapter 9 are almost at the opposite pole from a random phase assumption, and represent a deliberate overstatement, made in order to illustrate the theoretical situation as clearly as possible.

GENERAL FORMULATION OF THE DATA-SMOOTHING PROBLEM

ONE OF THE PRINCIPAL difficulties in any treatment of data smoothing is that of stating exactly what the problem is and what criteria should be applied in judging when we have a satisfactory solution. It is consequently necessary to embark upon a rather extensive general discussion of the data-smoothing problem before it is possible to consider specific methods of designing data-smoothing structures. This preliminary survey will occupy Chapters 7, 8, and 9. As a first step this chapter will describe two of the general ways in which the data-smoothing problem can be approached mathematically. The formulation of the problem which is finally reached in Chapter 9 is not the one which is most obviously suggested by these approaches. This, however, does not lessen their value in characterizing the problem broadly.

7.1 A PHYSICAL ILLUSTRATION

In an actual fire-control system the data-smoothing problem is usually made fairly specific because of the particular geometry adopted in the computer. It may be helpful to have some particular case in mind as a touchstone in interpreting the general discussion. For this purpose the most appropriate example is furnished by long range land-based antiaircraft fire, since most of the analysis described in this monograph was developed originally for its application to this problem. It is usually assumed in the antiaircraft problem that the target flies in a straight line at constant speed, and in one case at least the computer operates by converting the input data into Cartesian coordinates of target position and differentiating these to find the rates of travel in the several Cartesian directions. These rates form the basis of the extrapolation.

The process is illustrated in Figure 1. The input coordinates are transformed into electrical voltages proportional to x_p , y_p , and z_p , the Cartesian coordinates of present position,

in the coordinate converter at the left of the diagram. The extrapolation for x is shown explicitly. It consists essentially in differentiating to find the x component of target velocity, multiplying the derivative by the time of flight t_f , and adding the result to x_p to find

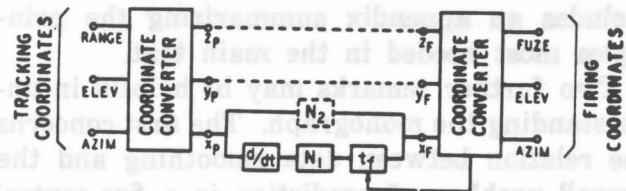


FIGURE 1. Data-smoothing networks in linear prediction circuit.

x_f , the predicted future value of x . A similar procedure fixes y_f and z_f . After the addition of certain ballistic corrections, these three coordinates of future position are transformed into gun aiming orders in the coordinate converter shown at the right of the drawing. This last unit also provides the time of flight required as a multiplier in the extrapolation.

The small irregularities in the input data caused by tracking errors are greatly magnified by the process of differentiation. It is thus necessary to smooth the rates considerably if a reliable extrapolation is to be secured. The data-smoothing network for the x coordinate is represented by N_1 in Figure 1. Since the Cartesian velocity components are theoretically constants if the assumption of a straight line course at constant speed is correct, a data-smoothing network in this computer must be essentially an averaging device which gives an appropriately weighted average of the fluctuating instantaneous rate values fed to it. The problem of "smoothing a constant" is given special attention in Chapter 10. Aside from the particular circuit of Figure 1, we may, of course, be required to smooth a constant whenever the prediction is based upon an assumed geometrical course involving one or more parameters which are isolated in the circuit.

In addition to smoothing the rates we can, if we like, attempt to smooth the irregularities in present position also. A network to accomplish this purpose is indicated by the broken line structure N_2 in Figure 1. Of course, in dealing with the present position we are no longer smoothing a constant, but suitable structures can be obtained by methods described later. However, the effect of tracking errors in the present position circuit is so much less than it is in the rate circuit that N_2 can generally be omitted.

Geometrical assumptions of the sort implied in Figure 1 are helpful in visualizing the problem, and they are of course of critical importance in determining what the final data-smoothing device will be. It is important not to make explicit assumptions of this kind too early in the formal analysis, however, since the meaning of such assumptions is one of the aspects of the general problem which must be investigated. For example, it is apparent that no airplane in fact flies exactly a straight line, nor flies a straight line for an indefinite period. In detail, the solution of the data-smoothing problem depends very largely on how we treat these departures from the idealized straight line path. For the present, consequently, it will be assumed that the input data are presented to the data-smoothing and predicting devices in terms of some generalized coordinates, the nature of which we will not inquire into too closely. A given coordinate might, for example, be a velocity, a radius of curvature, an angle of dive or climb, or any other quantity which would be directly useful in making a prediction, or it might be a simple position coordinate such as an azimuth or an altitude.

The data-smoothing and predicting operation itself is assumed to be performed by linear invariable devices. Aside from the fact that this assumption is, of course, a tremendously simplifying one, it also fits the data-smoothing problem very nicely, as the problem is formulated in this chapter. With other formulations, however, it appears that somewhat better results may be obtainable from variable devices or devices including more or less radical amounts of nonlinearity. These possibilities are discussed briefly in Chapter 14.

7.2

DATA SMOOTHING AND PREDICTION

Figure 1 illustrates a distinction between two possible methods of looking at the data-smoothing problem which it is advisable to establish for future purposes. In describing the x system in Figure 1 we laid emphasis on the particular networks N_1 and N_2 . It is clear, however, that the complete x circuit with input x_i and output x_r is a network having overall transmission properties which can be studied. Since t_i will normally vary with time, the network is not, strictly speaking, an invariable one, but the changes of t_i are ordinarily too slow to make this an essential consideration.

When it is necessary to make a distinction between these points of view, a network such as N_1 , which is merely an element in the prediction process, will be called a *data-smoothing structure*. An overall circuit, providing data smoothing and prediction in one step, will be called a *data-smoothing and prediction network*, or simply a *prediction network*. Although these points of view have been illustrated for rectangular coordinates, they obviously apply also in many other situations. For example, we might go so far as to apply the overall point of view to a complete circuit from input azimuth, say, to output azimuth.

Both points of view are taken from time to time in the monograph. When possible, however, principal attention has been given to the limited data-smoothing problem. This tends to simplify the discussion, since the limited problem is evidently more concrete than the overall prediction problem. Moreover, it permits us to deal lightly with such questions as the particular choice of coordinates in which the smoothing operations are conducted, since it assumes that the general kinematical framework of prediction has already been decided upon. On the other hand, the overall point of view is more effective in certain situations, and it is the only natural one in the statistical treatment described in the next section.

7.3

DATA SMOOTHING AS A PROBLEM IN TIME SERIES

The most direct and perhaps the most general approach to data smoothing consists in re-

garding it as a problem in time series. This is the approach used by Wiener in his well-known work.¹ It essentially classifies data smoothing and prediction as a branch of statistics. The input data, in other words, are thought of as constituting a series in time similar to weather records, stock market prices, production statistics, and the like. The well-developed tools of statistics for the interpretation and extrapolation of such series are thus made available for the data-smoothing and prediction problem.

To formulate the problem in these terms, let $f(t)$ represent the true value of one of the coordinates of the target and let $g(t)$ represent the observational error. Then $f(t)$ and $g(t)$ are both time series in the sense just defined. The set of all such functions corresponding to the various possible target courses and tracking errors form an ensemble of time series or a statistical population. One can imagine that a large number of particular functions $f(t)$ and $g(t)$ have been recorded, each with a frequency proportional to its actual frequency of occurrence. Wiener assumes that they are *stationary*, that is, that the statistical properties of the ensemble are independent of the origin of time. This, of course, implies that both functions exist from $t = -\infty$ to $t = +\infty$. We will sometimes find it more convenient to make the assumption that the two functions vanish after some fixed, but sufficiently remote, points on the positive and negative real t axis.*

The input signal to the computer is of course $f(t) + g(t)$. If we assume that the coordinate in question represents a position, the quantity we wish to obtain is $f(t + t_f)$, where t_f represents the prediction time. If the coordinate is a rate, we are interested in an average value of $f(t)$ over the prediction interval. This complicates the mathematics somewhat, but does not essentially affect the situation.

* This is done for technical mathematical reasons. We shall later have occasion to consider the Fourier transforms of $f(t)$ and $g(t)$, and, to have well-defined transforms, the integrals of the squares of the two functions, from $t = -\infty$ to $t = +\infty$, should be finite. This would not happen under the "stationary" assumption. Wiener avoids the difficulty by introducing what he calls a *generalized harmonic analysis*, but this method is far too complicated to be treated in a brief sketch like the present.

We shall not, of course, be able to predict $f(t + t_f)$ perfectly accurately. Let the predicted value be represented by $f^*(t + t_f)$. In virtue of our assumption that the data-smoothing and prediction circuit is to be a linear invariable network, the relation between $f^*(t + t_f)$ and the total input signal $f(t) + g(t)$ can be written as

$$f^*(t + t_f) = \int_{-\infty}^0 [f(\sigma) + g(\sigma)] dK(\sigma) \quad (1)$$

where $dK(\sigma)$ represents the effect of the data-smoothing and prediction circuit. Comparison to equations (2) and (5) of Appendix A shows that K is, in fact, the indicial admittance of this circuit. The particular problem to be solved is of course that of finding a shape for the function $K(\sigma)$ which will make $f^*(t + t_f)$ the best possible estimate of $f(t + t_f)$.

The fact that the upper limit of integration in equation (1) is taken as $\sigma = 0$ is particularly to be noted. It corresponds to the fact that in making a prediction we are entitled to use only the input data which has accumulated up to the prediction instant. This restriction will be conspicuous in the next chapter, where the time-series analysis is completed.

7.4 THE AUTOCORRELATION

The principal statistical tool used in studying equation (1) is the so-called autocorrelation. Under the "stationary" assumption the autocorrelation for $f(t)$ is defined by

$$\phi_1(\tau) = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T f(t + \tau) f(t) dt. \quad (2)$$

We can obtain a normalized autocorrelation, which is more convenient for some purposes, by dividing by $\phi_1(0)$. This gives

$$\phi_{1N}(\tau) = \frac{\phi_1(\tau)}{\phi_1(0)} = \lim_{T \rightarrow \infty} \frac{\int_{-T}^T f(t + \tau) f(t) dt}{\int_{-T}^T [f(t)]^2 dt}. \quad (3)$$

If we assume that $f(t)$ in fact vanishes for sufficiently large positive or negative values of t , the limit sign can be disregarded and $\phi_{1N}(\tau)$ becomes simply

$$\phi_{1,N}(\tau) = \frac{1}{W_1} \int_{-\infty}^{\infty} f(t + \tau) f(t) dt \quad (4)$$

where $W_1 = \int_{-\infty}^{\infty} [f(t)]^2 dt$ and represents the total "energy" in the time series $f(t)$.

Precisely similar expressions can be set up for the autocorrelation $\phi_2(\tau)$ or $\phi_{2,N}(\tau)$ of the observational error function $g(t)$. In a general case we might also have to worry about a possible cross correlation between $f(t)$ and $g(t)$. This would be represented by a cross-correlation function $\phi_{12}(\tau)$, obtained by integrating the product $f(t + \tau)g(t)$. In practical fire control, however, it can be assumed that the correlation between target course and tracking errors is small enough to be neglected.

As a simple example of the calculation of an autocorrelation we may assume that $f(t) = \sin \omega t$. Then

$$\begin{aligned} \phi_1(\tau) &= \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T \sin \omega(t + \tau) \sin \omega t \cdot dt \\ &= \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T \frac{1}{2} [\cos \omega \tau - \cos (2\omega t + \omega \tau)] dt \\ &= \frac{1}{2} \cos \omega \tau, \end{aligned} \quad (5)$$

since the term $\cos (2\omega t + \omega \tau)$ will contribute nothing in the limit.

The maximum value of $\phi_1(\tau)$ in (5) is found at $\tau = 0$. This is to be expected, since obviously the correlation between identical values of the function is the best possible. What is exceptional about the present result is the fact that $\phi_1(\tau)$ is not small for all large τ 's. This is fundamentally a consequence of the fact that we chose an analytic expression for $f(t)$, so that the relation between two values of the function is completely determinate, no matter how great the difference between their arguments. In a more representative time series, involving a certain amount of statistical uncertainty, we would expect $\phi_1(\tau)$ to approach zero as τ increases, reflecting the increasing importance of statistical dispersion as the time interval becomes greater.

The significance of the autocorrelation function for data smoothing and prediction is obvious without much study. Thus, suppose for

simplicity that the observational error $g(t)$ is zero. Then the autocorrelation $\phi_1(\tau)$ is the only one involved. It is a measure of the extent to which the true target path "hangs together" and is thus predictable. For example, in weather forecasting it is a well-known principle that in the absence of any other information it is a reasonably good bet that tomorrow's weather will be like today's but that the reliability of such a prediction diminishes rapidly if we attempt to go beyond two or three days. This would correspond to an autocorrelation function which is fairly large in the neighborhood of $\tau = 0$, but diminishes rapidly to zero thereafter.

In a similar way the autocorrelation of the observational error $g(t)$ represents the extent to which this error hangs together. In this case, however, a high correlation is exactly what we do not want. Thus, if $\phi_2(\tau)$ vanishes rapidly as τ increases from zero, closely neighboring values of g are quite uncorrelated, and we need only average the input data over a short interval in the immediate past in order to have most of the observational errors averaged out. If $\phi_2(\tau)$ is substantial for a much longer range, on the other hand, a much longer averaging period is necessary, with correspondingly greater uncertainties in the value obtained for $f(t)$.

7.5 THE LEAST SQUARES ASSUMPTION

The autocorrelation function does not in itself suffice to determine a time series completely. For example, it is easily seen that the functions $\sin t + \sin 2t$ and $\sin t + \cos 2t$ have the same autocorrelation in spite of the fact that they represent waves of quite different shape. The autocorrelation function, however, has a peculiar importance in the fact that under many circumstances it is the only piece of information about the time series which we need to know.

The significance of the autocorrelation becomes apparent as soon as we investigate the error in prediction. In many mathematical situations involving linear systems it is convenient to deal with the square of the error rather than with the error itself, since a first variation in the error squared expression gives a

linear relationship in the quantities of direct interest. We will deal with the square of the error here. If E represents the instantaneous error, $f^*(t + t_f) - f(t + t_f)$, the mean square error over a long period of time is evidently

$$\begin{aligned}\bar{E}^2 &= \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T [f^*(t + t_f) - f(t + t_f)]^2 dt \\ &= \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T [f(t + t_f)]^2 dt \\ &\quad - \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T}^T f(t + t_f) f^*(t + t_f) dt \\ &\quad + \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T [f^*(t + t_f)]^2 dt. \quad (6)\end{aligned}$$

The first integral in equation (6) can be evaluated immediately. From (2) it is $\phi_1(0)$. To evaluate the second integral replace $f^*(t + t_f)$ by its definition from (1). This gives

$$\begin{aligned}& - \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T}^T f(t + t_f) dt \int_0^\infty [f(t - \tau) \\ & \quad + g(t - \tau)] dK(\tau) = - \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^\infty dK(\tau) \\ & \quad \int_{-T}^T [f(t + t_f)f(t - \tau) + f(t + t_f)g(t - \tau)] dt\end{aligned}$$

if we reverse the order of integration. Since we assume that f and g are uncorrelated, however, the product $f(t + t_f)g(t - \tau)$ in this expression makes no contribution to the final result, and by replacing the integral of $f(t + t_f)f(t - \tau)$ by its value in terms of ϕ_1 , the expression as a whole can be written as

$$- 2 \int_0^\infty \phi_1(t_f + \tau) dK(\tau).$$

The third integral in (6) can be simplified in similar fashion. The final result becomes

$$\begin{aligned}\bar{E}^2 &= \phi_1(0) - 2 \int_0^\infty \phi_1(t_f + \tau) dK(\tau) \\ & \quad + \int_0^\infty dK(\tau) \int_0^\infty [\phi_1(\tau - \sigma) + \phi_2(\tau - \sigma)] dK(\sigma).\end{aligned} \quad (7)$$

The only quantities appearing in equation (7) are the autocorrelations, ϕ_1 and ϕ_2 , of the true target path and the observational error, and the function K which specifies the data-smoothing structure. The theoretical problem

with which we are confronted is evidently that of choosing K to make the mean square error as small as possible for any given ϕ 's. This problem will not be attacked here, although a solution obtained by a somewhat indirect method is presented in the next chapter. The principal reason for deriving equation (7) is to demonstrate the very important fact that *the mean square error depends only upon the two autocorrelations*. No other characteristics of the input data need be considered.

It will be recalled that the mean square criterion was introduced originally on the ground of mathematical convenience. This leaves unsettled the question of how good a measure of performance for a data-smoothing network it actually is. This is a critical question, since upon it depends the validity of the whole approach outlined in this chapter. A priori, the least squares criterion is a dubious one since it gives principal weight to large errors. In fire control we are normally interested only in shots which are close enough to register as hits. If a shot misses it makes little difference whether the miss is large or small. The merits of the least squares criterion are considered in more detail in Chapter 9, where the conclusion is reached that the criterion is probably adequate for many problems but needs to be supplemented or replaced in others, including the special case of heavy antiaircraft fire to which particular attention is given in this monograph. Pending the discussion in Chapter 9, the least squares criterion will be assumed to be a valid one, with the understanding that the analysis is intended primarily for its value in contributing to the general understanding of the data-smoothing problem rather than as a means of fixing the exact proportions of an optimal smoothing network.

7.6 DATA SMOOTHING AS A FILTER PROBLEM

The time-series approach to data smoothing is closely associated with another which at first sight may seem quite different. This second approach is suggested by the procedures used in communication engineering. Here the signals, be they voice, music, television, or what not, are again time series. Instead of dealing

with actual signals varying in a more or less irregular and random manner with time, however, it is customary to deal with their equivalent steady-state components on the frequency spectrum.^b

The analysis of data smoothing can conveniently be approached by supposing that both the true path of the target and the effects of tracking errors are represented, in a similar way, by their frequency spectra. When the situation is presented in this way, however, there is an obvious analogy between the problem of smoothing the data to eliminate or reduce the effect of tracking errors and the problem of separating a signal from interfering noise in communication systems. We may take as an example of the latter the transmission of voice or music by ordinary radio over fairly long distances, so that the effects of static interference are appreciable. In such a system a reasonable separation of the desired signal from the static can be obtained by means of a filter. In a representative situation an appropriate filter might transmit frequencies up to perhaps 2,000 or 3,000 cycles per second,^c while rejecting higher frequencies.

The choice of any specific cutoff, such as 2,000 or 3,000 c, in the radio system depends upon a compromise between conflicting considerations. Both speech or music and static normally include components of all frequencies which can be heard by the human ear. Thus, suppressing any frequency range below the limits of audibility, at perhaps 10,000 or 20,000 c, will injure the signal to some extent. The intensity of the signal components, however, diminishes rapidly above 2,000 or 3,000 c, while the energy of the static interference is more evenly distributed over the spectrum. Thus, by filtering out the first 2,000 or 3,000 c, we can retain most of the signal while rejecting most of the noise. Naturally, the exact dividing line will depend upon the relative levels of signal and noise power. If the static interference is quite weak, for example, it would be worth

while to transmit a considerably wider band in order to retain a more nearly perfect signal. If the static level is extremely high, on the other hand, it would be necessary to transmit a still narrower band at the cost of greater mutilation of the signal.

The separation of the true path of a target from the observed path including tracking errors, as a preliminary to prediction of the future position of the target, presents an approximately analogous situation. Again the spectrum of the "signal" or true path is concentrated principally in a low-frequency band, in most instances, while the energy of tracking errors or "noise" appears principally at considerably higher frequencies. Thus the two can be separated by a low-pass filter. The separation, however, is not complete since some components of the signal spectrum extend into the noise region. Thus the smoothing process must be accompanied by some mutilation of the signal, and the optimum compromise is again attained from a filter which transmits a relatively broad band when the tracking errors are of low intensity and a much narrower band when they are large.

In these terms the most obvious difference between the data-smoothing problem and the static interference problem in the radio system is in the order of magnitude of the frequencies involved. They are roughly 10,000 times smaller in the data-smoothing case. Thus, the typical signal band in a fire-control system may cover a few tenths of a cycle per second, in comparison with a useful band of 2,000 or 3,000 c in a radio system, and the spectrum of tracking errors or noise, with representative tracking devices, includes appreciable components up to perhaps 2 or 3 c, in comparison with a total effective noise band in the radio system extending to the limits of audibility at perhaps 20,000 c.

This analogy between data smoothing and the filtering problems which appear in ordinary communication systems transmitting speech or music must of course not be carried too far. For example, previous experience with communication filters is of no help in fixing in detail the cutoff in attenuation characteristic of the data-smoothing filter, since in communication systems these choices depend on psycho-

^b The review of communication theory given in Appendix A shows how this equivalence is established by Fourier or Laplace transform methods.

^c In practice, of course, the filtering would probably take place in the radio-frequency circuits, but it is more convenient here to think of it occurring in the demodulated output.

logical considerations of no relevance in the fire-control problem. Methods of determining the best rules for proportioning a data-smoothing filter, therefore, remain to be determined. We may also notice that, whereas the time-series approach was of the *data-smoothing and prediction* type, the filter approach emphasizes *data smoothing* only. The addition of the prediction function can be expected to change materially the overall characteristics of the circuit. Neither of these remarks, however, robs the filter approach of its value as a simple way of thinking about the problem qualitatively.

7.7 RELATION BETWEEN TIME-SERIES AND FILTER APPROACHES

The time-series and filter methods of looking at data smoothing are related to one another by the fact that the autocorrelation can be computed from the amplitude spectrum, or vice versa, by Fourier transform means. Consider, for example, the Fourier transform of the autocorrelation. If we make use in particular of (4) we have

$$\begin{aligned} & \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \phi_{1,N}(\tau) e^{-i\omega\tau} d\tau \\ &= \frac{1}{\sqrt{2\pi} W_1} \int_{-\infty}^{\infty} e^{-i\omega\tau} d\tau \int_{-\infty}^{\infty} f(t+\tau) f(t) dt \\ &= \frac{1}{\sqrt{2\pi} W_1} \int_{-\infty}^{\infty} f(t) dt \int_{-\infty}^{\infty} f(t+\tau) e^{-i\omega\tau} d\tau \\ &= \frac{1}{\sqrt{2\pi} W_1} \int_{-\infty}^{\infty} f(t) e^{i\omega t} dt \int_{-\infty}^{\infty} f(t+\tau) e^{-i\omega(t+\tau)} d\tau \\ &= \frac{\sqrt{2\pi}}{W_1} F(\omega) \bar{F}(\omega) \end{aligned} \quad (8)$$

where

$$\begin{aligned} F'(\omega) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t+\tau) e^{-i\omega(t+\tau)} d\tau. \end{aligned} \quad (9)$$

$F(\omega)$ is of course the steady-state spectrum of the signal $f(t)$. Equation (8) thus states that the Fourier transform of $\phi_{1,N}$ is equal to a constant times the square of the amplitude of the steady-state spectrum. The amplitude squared spectrum is, however, a measure of

the power per cycle. The relation is therefore equivalent to the statement that the autocorrelation and power spectrum are Fourier transforms of each other.

Since we have already established the fact that the mean square error in prediction depends only on the autocorrelation, this analysis enables us to conclude immediately that the mean square error can also be calculated from the power spectra of the signal and noise. It is entirely independent of the phase relations in either signal or noise. The phase characteristics of the data-smoothing network, which operates on the signal after a specific wave shape has been established, is, of course, still of consequence.

7.8 PHYSICAL AND TACTICAL CONSIDERATIONS

Thus far the material which has been presented has been primarily mathematical. It has consisted, in other words, of outlines of general analytical methods which are available for use with the data-smoothing problem. It is also possible to approach the problem in a much more concrete fashion. It is obvious that by giving thought to the details of the physical characteristics of tracking units and targets, and to the tactical situations with which we expect to deal, it should be possible to draw a number of specific conclusions about the problem as a whole. In a general theory of the design and tactical use of fire-control apparatus such an approach might well be a primary one. It is scarcely possible to follow it in detail in the present discussion. The following paragraphs, however, indicate some of the kinds of considerations which can be brought into the problem in this way. It will be seen that they tend to modify the strictly mathematical approach, partly by qualifying to some extent the assumptions made in the mathematics, and partly by tending to give much more emphasis to particular aspects of the problem than would appear in a general analytic outline.

CHOICE OF COORDINATES

One of the most obvious omissions in the general analysis thus far is any consideration of the choice of coordinates in which the data

smoothing is to take place. So far as either the statistical or filter theory is concerned, the coordinates in the data smoother may represent either the original tracking data or any transformation of them. The fact that there is actually something to be decided here, however, is easily seen from the long-range antiaircraft problem. The input tracking coordinates for antiaircraft would normally be azimuth, elevation, and slant range. If the airplane flies in a straight line roughly overhead, the general shape of the azimuth and the azimuth rate as functions of time are given by the curves in Figure 2. The curves become indefinitely

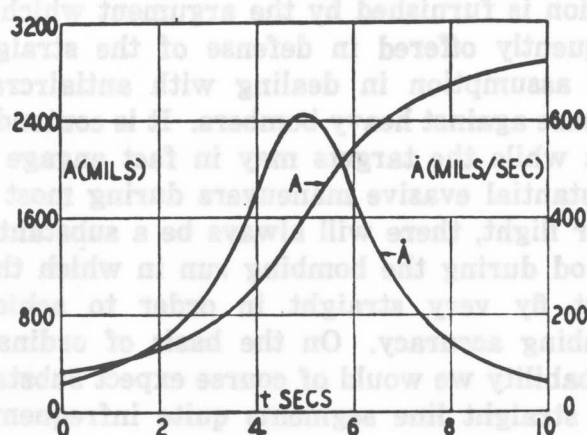


FIGURE 2. Azimuth and azimuth rate for crossing target.

steeper as the target path approaches the zenith, and it will be seen that if the approach is reasonably close, either the azimuth or the azimuth rate must include a very substantial amount of high-frequency energy. Since the possibility of an effective separation between the signal and noise in the filter approach depends upon the assumption that the signal components are of quite low frequency with respect to the noise, the presence of this high-frequency energy is evidently serious.

When the target describes a violently evasive path the signal spectrum must naturally include substantial high-frequency components, whatever the coordinate system may be. The high-frequency components indicated in Figure 2, however, are due to the fact that the target path happens to pass almost over the director and are essentially superimposed upon the high-frequency components which reflect the complexity of the target path itself. It is clear

as a matter of principle that an acceptable coordinate system for data smoothing should not introduce frequency components which depend upon such accidental factors as the location and orientation of the coordinate system. The rectangular system mentioned in connection with Figure 1 evidently meets this condition; so also does the "intrinsic" system described in the next section.

PHYSICAL LIMITATIONS OF TARGET OR TRACKER

We may also approach the data-smoothing question by a consideration of the motions which are physically possible either in the target or in the tracking device. In the heavy antiaircraft problem, for example, there are substantial physical limitations on the performance possibilities of present-day aircraft. We can be quite sure that any motion incompatible with these limitations is necessarily a tracking error and can be removed from the incoming data. Naturally, these limitations must appear in the power spectrum of the signal if they affect the mean square error in prediction, so that their existence in no way disputes the mathematical framework we have set up. Consideration of the physical factors which produce them, however, may permit them to be established more easily or in more clear-cut fashion than would be possible from a statistical examination of target records alone.

The limitations on airplane performance can be stated most simply when the motion of the airplane is expressed in so-called intrinsic coordinates. These are the speed of the airplane, its heading, and its angle of dive or climb. The maneuvering possibilities of a conventional airplane in these three directions are quite unequal. By banking sharply it can maneuver violently to the right and left and thus make quick changes in heading. The possibilities of maneuvering up and down, however, are considerably less, particularly for a heavy airplane, where there are usually restrictions on the maximum angle of dive or climb which can be assumed. The possibilities of quickly changing the speed of the airplane, finally, are almost nil. The thrust of an airplane propeller is so small in comparison with

the mass of the airplane that only small accelerations are possible.⁴

Thus the optimum filters for the three coordinates should be different. The one for speed can have a very narrow band, since most of the signal energy for this coordinate occurs at very low frequencies. The optimum band for the angle of dive or climb, however, should be larger (unless it turns out that pilots seldom make use of maneuvering possibilities in this direction) and the one for the heading larger still. In this ability to discriminate among the various possible directions of motion the intrinsic coordinate system is evidently an improvement even on the rectangular system.

SETTLING TIME

Another aspect of the data-smoothing problem which has not been given conspicuous attention in the purely mathematical discussion is the fact that in an actual tactical situation questions of elapsed time are of great importance. Engagements usually begin suddenly and last for a comparatively brief period, and it is important to find a data-smoothing scheme which provides adequate firing data as quickly as possible after an engagement starts. A situation essentially similar to the beginning of an engagement may also be presented whenever the target makes a sudden change of course or whenever it is necessary to shift from one target to another in a given attacking body. The time required for a computer to give usable output data after any of these events is its so-called "settling time," and is one of the most important parameters of any data-smoothing system. It is possible to make rough estimates of settling time by indirect means in both the statistical and filter theories of data smoothing, but no explicit consideration of necessary time lapses appears in either theory. Evidently, the fundamental fault lies with the "stationary" assumption.

⁴ This ignores the possibility of changing the speed through gravitational forces. Since these possibilities are linked to the angle of dive or climb, however, they can be predicted. This has actually been done in one experimental computer.

EFFECT OF HUMAN FACTORS

Aside from the conditions on target performance which arise from the physical characteristics of the target itself, there are others which are due to the fact that the target is under the control of a human being with a definite purpose. The language of the statistical and filter methods is broad enough to cover almost any situation. It tends to suggest, however, that the typical target paths with which we deal are the relatively structureless consequences of random physical forces. The intervention of purposive human behavior, on the other hand, tends to give paths which fall into more or less definite patterns. A simple illustration is furnished by the argument which is frequently offered in defense of the straight line assumption in dealing with antiaircraft defense against heavy bombers. It is contended that while the targets may in fact engage in substantial evasive maneuvers during most of their flight, there will always be a substantial period during the bombing run in which they must fly very straight in order to achieve bombing accuracy. On the basis of ordinary probability we would of course expect substantial straight line segments quite infrequently if the course as a whole shows marked dispersion, and the intervention of the human pilot thus provides a higher degree of structure than one would expect in a corresponding situation dominated by purely natural factors.

A broader example is furnished by a comparison of two airplanes, or perhaps more simply of two boats, one of which is under the control of a human operator, while in the other the steering controls are lashed in a neutral position. Both boats, say, may be expected to experience small variations of course due to the random effects of wind and waves upon them. Over a short period of time the observed motions of the two boats should be substantially identical. In the case of the boat with the lashed helm these random variations will tend to accumulate, so that it is possible to make a reasonable prediction of the position of the boat for only a comparatively short distance in the future. In the boat with the human steersman, on the other hand, we may expect corrections to be applied as soon as the random effects become large, so that the boat tends to

retain the same general course and it is possible to predict its position hours or even days later from a relatively brief observation.

Neither of these illustrations is inconsistent with the mathematical framework laid down earlier in the chapter, in a purely theoretical sense. For example, the bombing run illustration merely states that because of the presence of the human operator there are definite phase relations in the input signal. As we have seen, such relations can exist without affecting computations based on mean square error. The comparison between the piloted and pilotless boats can be interpreted as the result primarily of differences in the signal power spectrum. In the case of the pilotless boat, for example, the signal occupies a fairly continuous low-frequency band, while in the case of the piloted boat it must be regarded as concentrated very closely around zero frequency, so that it is approximately a line spectrum superimposed on a continuous one. The formal mathematical theory covers also such cases as these.

The point of this discussion, however, is that the mathematical theory, although it is sufficiently general in a formal sense, fails to differentiate between such situations as those just described and the more shapeless sort involving continuous spectra with random

phase relations, even if the special features in these situations may be the controlling factors in determining the actual probability of hitting. If we could believe the bombing run hypothesis, for example, and had a sufficiently accurate computer and gun, we could expect to score a hit in every engagement, no matter how large the mean square error might be. More generally, it is probably only the tendency of targets to exhibit "line spectra" which prevents the real probability of a kill, small at best, from becoming microscopic. It is necessary to lay special emphasis on these factors in order to keep the overall fire control picture in perspective.

CRITERION OF PERFORMANCE

Last on this list of doubts about the statistical and filter theories, we may mention the least squares criterion of accuracy. This was discussed before, but it is mentioned again as a matter of emphasis, and because of its close relation with the factors we have just discussed. For example, the bombing run illustration obviously represents one situation in which the mean square error is not a good guide to the actual probability of scoring a hit.

STEADY-STATE ANALYSIS OF DATA SMOOTHING

IT WAS SHOWN in the previous chapter that both the statistical and filter theory ways of looking at the data-smoothing problem lead naturally to an analysis in terms of the power spectra of the signal and noise. The phase relations are not important as long as we accept the mean square error as a criterion of performance. The inadequacies of the mean square criterion will finally force us to abandon the steady-state attack in favor of a direct analysis in terms of the wave shapes of some assumed signals. The steady-state attack is nevertheless a very useful one. This chapter will consequently continue the analysis from this point of view. It will be assumed as heretofore that the heavy anti-aircraft problem is the particular subject of interest.

A large part of the discussion hinges upon the conditions which must be satisfied by the external characteristics of an electrical network if it is to be capable of physical realization in any way whatever. These limitations and the characteristics which may be postulated for physical networks are decisive since, in the absence of such restrictions, no limits could be set upon the performance which might be expected from data-smoothing and predicting circuits. The facts about physically realizable networks which we shall find of most use are summarized below, but the reader not familiar with this field is urged to read also the account given in Sections A.9 and A.10, Appendix A.^{15a}

The conditions which must be satisfied by physically realizable networks can be stated in either transient or steady-state terms. In transient terms they are expressed most simply by the statement that the response of a physical network to an impulsive force must be zero up to the time the force is applied. Thus the network has no power to predict a purely arbitrary event. That is, it has no way of foreseeing whether or not an impulse is actually going to be applied to it. This characteristic of physical networks is taken as a postulate.

The steady-state limitations on physical net-

works are expressed in terms of their attenuation and phase characteristics. They may be derived either from the transient specification or from the postulate that a physical network must be stable. There are no important limitations to be placed upon the attenuation and phase characteristics of physical networks as long as we deal with these characteristics separately, but there are very severe limitations on the phase characteristic which can be associated with any given attenuation characteristic or vice versa. In particular, when the attenuation characteristic is prescribed, there is a definite formula for calculating the unique limiting phase characteristic with which it may be associated.^{15b} This is the so-called "minimum phase" characteristic because any other physical network having the postulated attenuation characteristic must have as great or greater phase shift at every frequency. As we shall see later, this greater phase characteristic would correspond to longer lags in obtaining usable data, so that the minimum phase characteristic is the optimum for a data-smoothing network. The minimum phase characteristic has the additional important property that not only does it specify the transfer admittance of a physical network, but the reciprocal of that transfer admittance can also be realized by a physical structure.*

In addition to this principal formula for the relation between attenuation and phase there are a number of subsidiary expressions for special aspects of the problem. One in particular, relating the attenuation to the behavior of the phase characteristic in the neighborhood of zero frequency, is used extensively in this chapter.

* In limiting cases, such as may be found when the transfer admittance contains zeros or poles exactly on the real frequency axis, the "physical structure" may require such constituents as ideally nondissipative reactances, perfect amplifiers with unlimited gain, etc. This, however, is of no consequence for the present general discussion.

8.1 THE SIGNAL SPECTRUM

It is natural to begin with a discussion of the spectrum of a typical target path. Unfortunately no data on the spectra of actual measured airplane paths exist, and the theoretical assumptions which may be made about paths of airplane targets are best discussed in the next chapter. This section consequently will be confined to rather general observations about the problem. It will be convenient to assume for definiteness that the quantities to be smoothed are the velocity components in Cartesian coordinates.

The simplest point of departure is furnished by the conventional assumption that the target flies in a straight line at constant speed. If we could construe this assumption literally, it would mean that the velocity spectrum in rectangular coordinates would reduce to a single line at zero frequency. In practice, of course, the spectrum is not so simple. Even in the absence of deliberate maneuvering, the target will fly a slightly curved path because of "wander." Moreover, even if the target could fly exactly straight, the single line spectrum would apply only to a straight course indefinitely continued. The spectrum becomes more complicated if we consider the fact that tracking must have begun at some finite time in the past, or that the target may presumably change occasionally from one straight line course to another.

As a result of both these causes, the actual signal spectrum must be regarded as occupying a band bordering on zero frequency. The distribution of energy in detail will, of course, depend on particular circumstances. The band has no very well defined upper limit, but in most cases the great bulk, at least, of the energy should be below, say, one-fourth or one-fifth of a cycle per second. For example, the natural periods of a heavy airplane, which one would expect to be correlated with wander, are below this limit.¹ This limit is also sufficient to include most of the energy resulting from changes in course occurring as frequently as every ten or twenty seconds.

In general, it is to be supposed that the signal spectrum varies as ω^{-n} , where n may be 1, 2, 3, depending on the frequency range. This follows from general considerations of the

limitations of airplane performance. Thus, if we suppose that the velocity changes discontinuously from time to time, it follows from general Fourier principles that the amplitude must vary as ω^{-1} . This is presumably a fair representation of the actual signal spectrum at low frequencies. At moderate frequencies, however, we must take account of the fact that the velocity can actually be changed rapidly but not discontinuously, and we consequently assume that the amplitude begins to vary as ω^{-2} . Finally, at frequencies of the order of perhaps one cycle per second one must take account of the fact that the airplane must bank in order to turn. Since it takes some time to roll into the bank, even the acceleration in the lateral direction cannot be discontinuous, and consequently the amplitude must begin to vary as ω^{-3} . The application of such successive limiting factors in constructing a complete spectrum is described in more detail in Section A.8 of Appendix A.

One other general condition of the same kind can be mentioned. It can be shown² that the integral from zero to infinity of $\log H/1 + \omega^{-2}$, where H is the power spectrum, is very important in determining the properties of a time series. More explicitly, the integral converges if the series is essentially statistical, so that we cannot foretell the future from the past with absolute certainty. This of course is the case with an actual signal spectrum in a fire-control problem. It implies two consequences; first, that H cannot be zero over any finite band; and second, that in the neighborhood of infinite frequency H diminishes slowly enough so that $|\log H|/\omega \rightarrow 0$.

8.2 THE NOISE SPECTRUM

The spectrum of tracking errors depends largely upon the particular sort of tracking equipment involved. Broadly speaking, optical tracking equipment (at least that of the present or recent past) tends to produce tracking errors not only of small amplitude, but also of low frequency, so that they are hard to separate from the signal spectrum. Radar equipment, of the present time, produces higher-frequency errors. Relatively high-frequency errors are particularly likely to be found in very stiff automatic tracking radars.

A number of examples of spectra of tracking errors are shown in Figures 1, 2, and 3. The spectra are given directly in terms of range and angle errors. To make them comparable with the velocity spectra described previously

"random noise" functions.^b A random noise can be defined as a function which has a definite amplitude spectrum but completely random phase characteristics. The theory of such functions is well developed because of their frequent

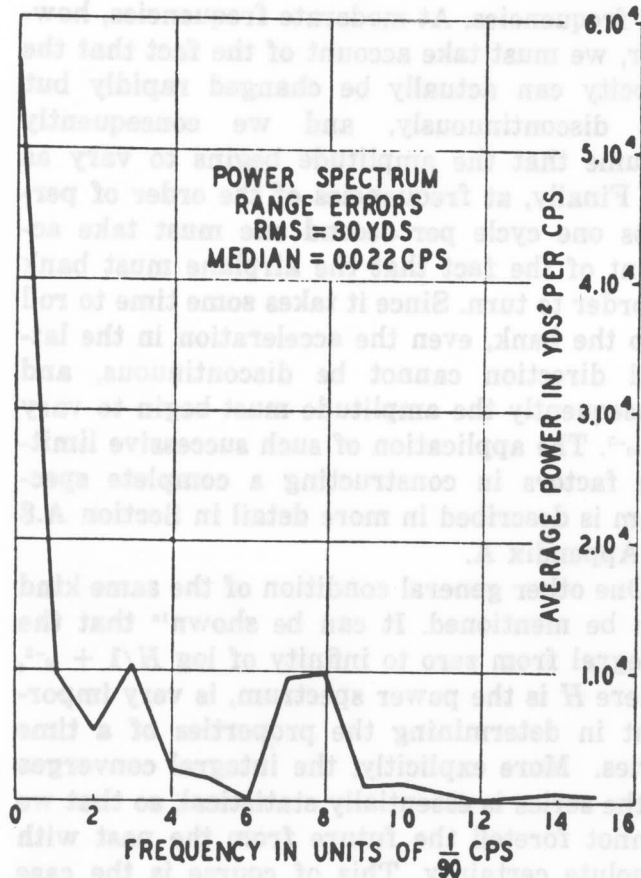


FIGURE 1. Power spectrum of range errors of experimental radar.

it would be necessary to multiply all amplitudes by ω . In addition, it would of course also be necessary to multiply the angle rates by some suitable range in order to compare them directly with the yards-per-second rates we have otherwise considered.

After multiplication by ω , the radar spectra appear to be about flat up to perhaps one cycle. Beyond that point they no doubt drop off slowly, although the accuracy of the data is not sufficient to permit the situation to be stated very exactly.

5.3 RANDOM NOISE FUNCTIONS

The properties of the signal and noise as we shall assume them here can be conveniently expressed by reference to the theory of so-called

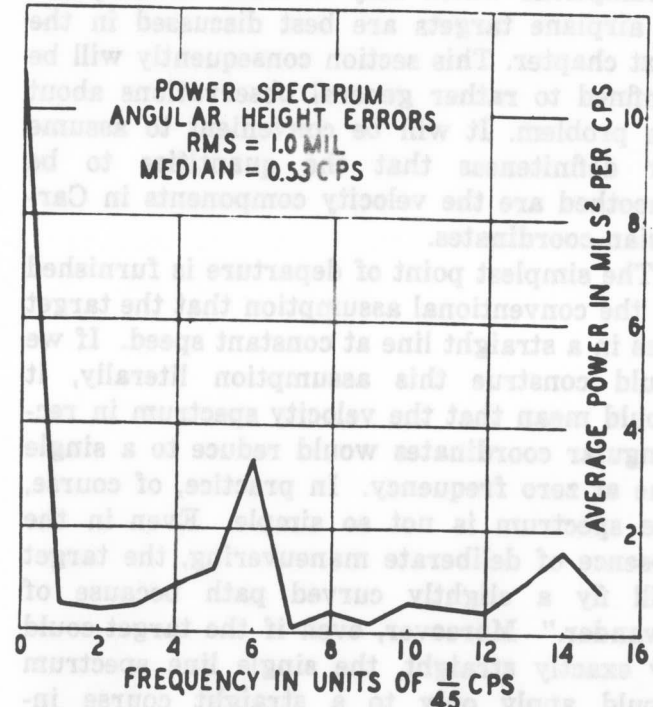


FIGURE 2. Power spectrum of angular height errors of experimental radar.

occurrence in physics. It is probable that neither our noise functions nor our signal functions are, strictly speaking, random noise according to this definition. Thus, there are probably certain definite phase relations in our noise functions because of the physical characteristics of tracking devices. There is no evidence, however, that any such relations are important enough to be significant in the data-smoothing problem, so that we are fully justified in identifying them with random noise functions as defined above. The phase relations in the signal are by no means random. As long as we consider only the mean square error, however, this factor is immaterial, and we can replace the actual signal by a random noise function with the same power spectrum for purposes of analysis.

The most familiar example of a random noise function is furnished by the thermal

^b The fact that we also refer to tracking errors as "noise" is, of course, merely a coincidence.

voltage across a resistance R . This is a random noise whose spectrum is constant up to very high frequencies with the value $P = 4kTR$ (k is Boltzmann's constant and T the absolute temperature). A second example is black body

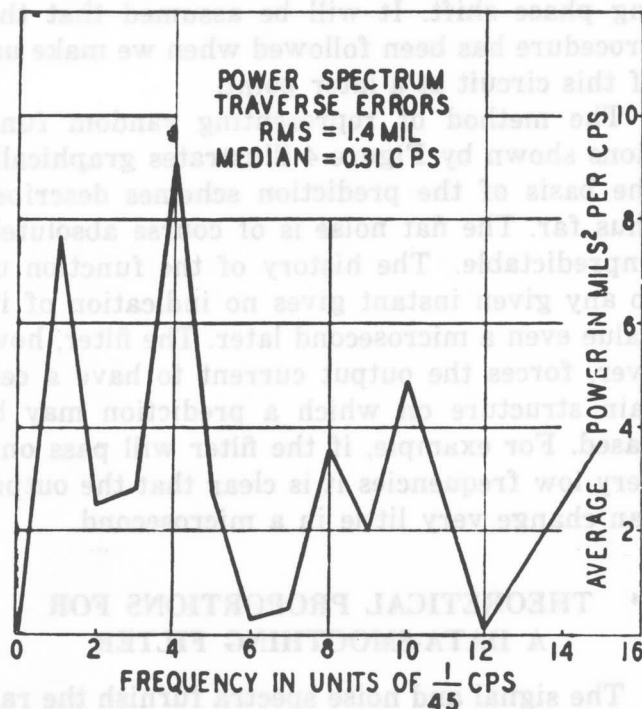


FIGURE 3. Power spectrum of traverse angular errors of experimental radar.

radiation. If there is black body radiation in a space, the electric (or magnetic) field intensity at a point is a random noise function with spectrum

$$P(f) = \frac{8\pi f^3}{C^3} \frac{1}{e^{hf/kt} - 1}$$

according to Planck's law. Random noise functions also occur in the Schottky effect, in Brownian motion, and in diffusion and heat flow problems.

For purposes of analysis, a random noise function can be thought of as a function made up of a large number of sinusoidal components, which are very closely spaced in frequency and whose phases are completely random.^{21-23a} Thus a random noise can be represented as

$$\sum_{n=1}^N a_n \cos(\omega_n t + \phi_n)$$

where $\omega_n = n\Delta f$, Δf being the frequency difference between adjacent components. The phase

angles ϕ_n are random variables which are independent with a uniform probability distribution from 0 to 2π . As Δf decreases the functions in this ensemble approach, in a certain sense, a limiting ensemble, providing the amplitudes a_n are adjusted properly. What is desired is to have the total power in the neighborhood of each frequency approach a certain limit $P(f)$, the power spectrum at that frequency. To do this we make

$$a_n^2 = 2\pi P(f)\Delta f.$$

In the limiting ensemble the total power within a small frequency range Δf is then $P(f)\Delta f$. The function $P(\omega)$ completely describes the random noise ensemble from the statistical point of view.

A particularly important special case is that of a random noise with a constant power spectrum. This is often called "flat" or "white" noise. True constancy out to infinite frequencies is of course impossible since it would imply an infinite total power in the function. The idea is, however, still useful and can be approximated, as with resistance noise, by having a spectrum which is constant out to such high frequencies that behavior beyond this point is of no importance to the problem. We may conveniently think of flat random noise as being made up of a succession of weak impulses occurring frequently but at random times with respect to one another. This results from the fact that a Fourier analysis of a single impulse gives a flat spectrum, and the random occurrence of many of them produces a random set of phases. In a physical problem, such as resistance noise or Brownian motion, these impulses might correspond to the effects of individual small particles. Such a situation is of course completely chaotic. If the impulses are large and occur relatively infrequently, the power spectrum is still flat, though the function is no longer a random noise function as defined here. This conception, which corresponds to a physical situation including definite causative elements, will be revived later under the name of the elementary pulse method of analysis.

Random noise functions have a number of interesting characteristics. For example, they have the "ergodic property." This means that

averaging a statistic along the length of a particular random function gives the same results as averaging the same statistic over an ensemble of functions having the same power spectrum. Each function is typical of the ensemble. To be more precise one must admit exceptions, but the probability of an exception is zero. For example, if we determine the fraction of time a given random function $f(t)$ has a value greater than some constant A , it will be equal to the fraction of all functions in the ensemble which are greater than A at $t = 0$ (with probability 1).

A second characteristic of random noise functions is the fact that they frequently lead to Gaussian or normal law distributions. For example, the amplitudes of a random noise function are distributed about zero in accordance with the normal error law. Likewise, the amplitudes for two points spaced a given distance apart form a two-dimensional normal error law distribution when we consider all possible positions of the first point. It is apparent that if the signal and noise are actually random functions the mean square error is as good a criterion of performance as any other, since it completely fixes the distribution in a normal law case.

A final property of random noise functions is the fact that if a random noise is passed through a filter the output is still a random noise. If the power spectrum of the noise is $P(\omega)$ and the transfer characteristic of the filter is $Y(i\omega)$, the output spectrum is $P(\omega)|Y(i\omega)|^2$. In particular, if we take the derivative of a random noise with spectrum $P(\omega)$ we obtain one with spectrum $\omega^2 P(\omega)$.

This last property of random noise functions suggests a method of representing them which we shall find useful in the future. The method is represented by Figure 4. It consists of a

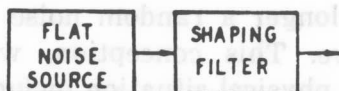


FIGURE 4. Circuit representation of random functions.

source of flat noise followed by a shaping filter to give the desired power spectrum. We can easily assign to the filter the characteristics of a physically realizable structure by making use

of the relations between attenuation and phase mentioned earlier in the chapter. It is merely necessary to convert the desired power spectrum into a specification of the attenuation characteristic of the filter and then use the loss-phase formula to compute the corresponding phase shift. It will be assumed that this procedure has been followed when we make use of this circuit at a later point.

The method of representing random functions shown by Figure 4 illustrates graphically the basis of the prediction schemes described thus far. The flat noise is of course absolutely unpredictable. The history of the function up to any given instant gives no indication of its value even a microsecond later. The filter, however, forces the output current to have a certain structure on which a prediction may be based. For example, if the filter will pass only very low frequencies it is clear that the output can change very little in a microsecond.

8.4 THEORETICAL PROPORTIONS FOR A DATA-SMOOTHING FILTER

The signal and noise spectra furnish the raw material from which a suitable data-smoothing filter can be deduced. We have still to determine, however, the exact rule for choosing the cutoff and attenuation characteristic of the filter from these spectra. It is clear that previous experience with signal-to-noise problems in systems transmitting voice or music is no help, since the filter proportions here depend upon psychological considerations of no relevance to the fire-control problem. For example, the interfering effect of a small amount of noise is much greater than one might expect from energy considerations, especially in intervals of low message level, and it is consequently worth while to maintain a relatively high level of attenuation in the noise band. Conversely, the breadth of the band required for the message depends as much on the ability of the ear to reconstruct a complete signal from an incomplete one as it does upon the actual signal power spectrum.

In the data-smoothing case a suitable criterion, dependent upon more physical considerations, can be obtained by minimizing the rms error at the filter output. This criterion is

easily developed from the power spectrum approach, and in a sense it is, of course, the only possible one as long as we follow the methods developed thus far.

A very general theory for the minimization of the rms error of the filter output has been developed by Wiener.¹ Since the power spectrum approach is not the one we shall eventually follow, however, it is not necessary to give this analysis in detail. The nature of the relationships can be seen from an elementary computation. Thus in Figure 5 let OA be a unit

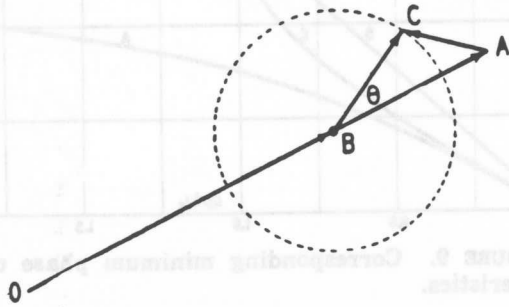


FIGURE 5. Vector relation between input and output of data-smoothing network.

vector representing the signal component at some particular frequency. Let the amplitude ratio between the input and output of the data-smoothing filter be x , and let it be assumed that the system is phase distortionless. This can always be accomplished, at the cost of lag, by phase equalization. Then the actual signal output can be represented by OB , where $OB/OA = x$. Let the ratio of noise power to signal power at this frequency be k^2 . Then the output noise can be represented by the vector BC , at some arbitrary phase angle θ , where $BC/OA = kx$.

The error in the output of the data-smoothing filter is evidently represented by the vector AC . We have

$$(AC)^2 = (OA)^2 [(1 - x - kx \cos \theta)^2 + (kx \sin \theta)^2] \\ = (OA)^2 [(1 - x^2) - 2kx(1 - x) \cos \theta + k^2 x^2].$$

Since θ is random the cross-product term involving $\cos \theta$ disappears on the average. (More generally, it disappears as long as the noise and signal are uncorrelated, whether or not their relative phases are entirely random.) This leaves the mean square error as

$$(AC)^2_{\text{mean}} = (OA)^2 [1 - 2x + (1 + k^2)x^2]. \quad (1)$$

The mean square error is a minimum if

$$x = \frac{1}{1 + k^2} = \frac{P_s}{P_N + P_s}$$

where P_s and P_N are, respectively, the signal and noise power at this frequency. Upon substituting this result in equation (1) and remembering that $(OA)^2 = P_s$, we find that the minimum mean square error is

$$(AC)^2_{\text{mean min}} = \frac{P_N P_s}{P_N + P_s}. \quad (2)$$

Equation (2) evidently represents the sought-for rule for the filter transmission characteristic. It is illustrated in Figure 6, where P_N

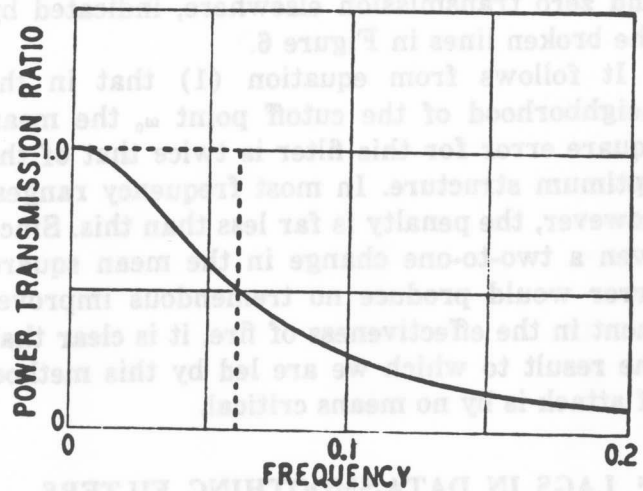


FIGURE 6. Optimum transmission characteristic for data smoothing assuming signals with random noise characteristics.

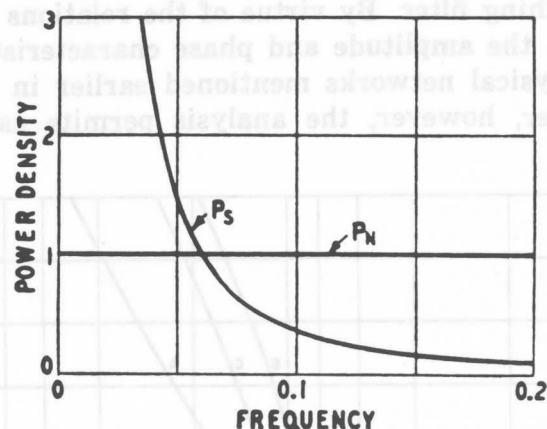


FIGURE 7. Signal and noise power spectra assumed in Figure 6.

and P_s have been chosen respectively as the flat curve and the $1/\omega^2$ curve in Figure 7. In comparison with the characteristics of typical filters in communication systems it is quite

rounded with a relatively slowly falling amplitude characteristic. More important than the detailed rule for the transmission characteristic, however, is the conclusion that the shape of the characteristic is not very critical. There is very little loss in replacing the actual curve in Figure 6, by any other similar characteristic. For example, we might validate the assumption of zero phase distortion by making use of the curve which automatically gives a linear phase shift.^{15c}

A more extreme illustration is furnished by the infinitely selective filter characteristic, with perfect transmission in the range in which the signal power is greater than the noise power, and zero transmission elsewhere, indicated by the broken lines in Figure 6.

It follows from equation (1) that in the neighborhood of the cutoff point ω_0 the mean square error for this filter is twice that of the optimum structure. In most frequency ranges, however, the penalty is far less than this. Since even a two-to-one change in the mean square error would produce no tremendous improvement in the effectiveness of fire, it is clear that the result to which we are led by this method of attack is by no means critical.

8.5 LAGS IN DATA-SMOOTHING FILTERS

The analysis just concluded has been directed at the amplitude characteristics of a data-smoothing filter. By virtue of the relations between the amplitude and phase characteristics of physical networks mentioned earlier in the chapter, however, the analysis permits us to

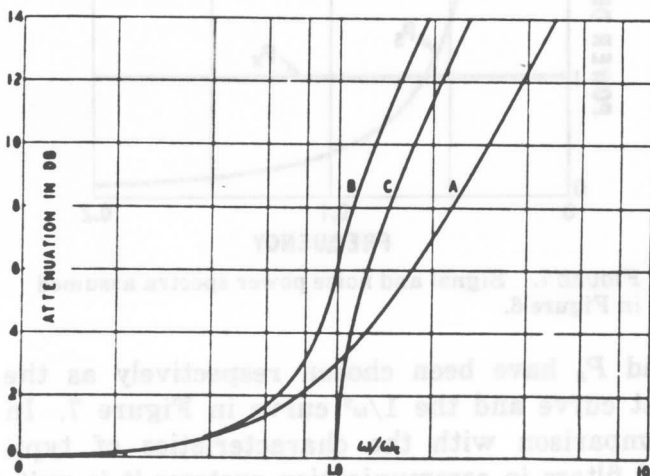


FIGURE 8. Some filter attenuation characteristics.

give at least a partial description also of the phase characteristics of the filters. This is an important consideration because it bears upon the question of time delays in data-smoothing systems which was mentioned in Chapter 7.

The general nature of the relationship in simple cases is illustrated by Figures 8 and 9.

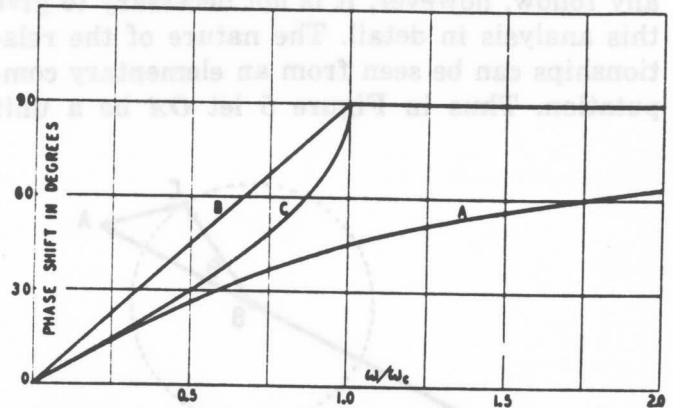


FIGURE 9. Corresponding minimum phase characteristics.

Figure 8 shows a series of rising attenuation characteristics equivalent to rather unselective falling amplitude characteristics of the general type shown by the principal curve in Figure 6. Figure 9 shows the corresponding phase characteristics computed on a minimum phase shift basis. In Figure 8 the central attenuation characteristic *B* has been so chosen that the corresponding phase characteristic in Figure 9 is exactly a straight line at low frequencies, where the transmitted amplitudes are appreciable. Curves *A* and *C* in the two drawings show slightly different cases, but it is clear from the figures that the tendency of the phase characteristics to approximate linearity is still marked.

In communication engineering a phase characteristic proportional to frequency is interpreted as indicating a delay in seconds equal to the slope $dB/d\omega$ of the phase characteristic. This relation is illustrated most simply by an ideal line. The ideal line has zero attenuation combined with a phase shift which is proportional to frequency and which at any given frequency is also proportional to the length of the line in question. If we apply any arbitrary wave to the line it is propagated down the line with a definite velocity and unchanged wave form. The time required for the wave to reach

any point on the line is equal to the slope of the phase characteristic to that point.

In a structure like a filter, which has an attenuation characteristic varying with frequency, it is of course no longer possible to transmit an arbitrarily impressed wave without change in wave shape. Even if the applied wave is merely a suddenly applied d-c voltage or single frequency sinusoid, there is a transient period before the response approximates its final value. In structures having a substantially linear phase characteristic over any frequency range in which they exhibit an appreciable amplitude response, however, this total transient characteristic falls naturally into two parts. The first is a waiting period equal to the slope of the phase characteristic, during which the response is very small, whereas the second is a true transient period in which the response is substantial but does not resemble the final steady-state response. This is illustrated by Figure 10 which shows the voltage at the fifth

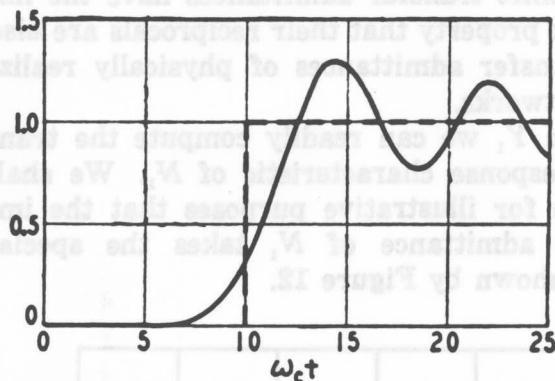


FIGURE 10. Voltage at fifth section of conventional low-pass filter in response to unit d-c voltage.

section of a conventional low-pass filter in response to a d-c voltage applied at zero time at the input terminals.¹⁰ The end of the waiting period, as deduced from the slope of the phase characteristic, is indicated by the broken line.

Delays of the sort just illustrated must be expected in a data-smoothing filter whenever the nature of the signal is changed. This happens at the beginning of tracking, in changing from one target to another, or even in following a single target when the target makes an abrupt change in course. Since usable data in a fire-control system must be quite accurate, the delay to be allowed for must include both the initial waiting period and the subsequent

transient period until the transient ripples have almost vanished. A considerable part of the art of designing data-smoothing networks consists in controlling the design so that these final transient ripples decay relatively rapidly. We are not yet ready to discuss this problem. It will turn out, however, that the minimum interval which can be assigned to the "true transient" period is about equal to that which must be allowed for the initial waiting period.^c Thus the slope of the phase characteristic can be used as an index of the lags which must be expected in data smoothing merely by doubling the delay to which the slope would normally be said to correspond.

When we use the phase slope as an index of delay it becomes immediately apparent that lags are the necessary consequence of smoothing in physical circuits. This is easily seen by reference to the relations which must exist between attenuation and phase characteristics in physical structures. An example is provided by the formula^{15d}

$$\left(\frac{dB}{d\omega}\right)_{\omega=0} = \frac{2}{\pi} \int_0^{\infty} \frac{A - A_0}{\omega^2} d\omega = \frac{2}{\pi} \int_0^{\infty} (A - A_0) d\left(\frac{1}{\omega}\right) \quad (3)$$

where A is attenuation, A_0 is the attenuation at zero frequency, and B is phase shift. In other words, the delay (measured by the slope of the phase characteristic at zero frequency) is proportional to the integral of the attenuation on an inverse frequency scale when the attenuation at zero frequency is taken as the reference. The equation thus states that the system will exhibit a lagging response as long as there is a net high-frequency attenuation. As a numerical illustration, let it be supposed that A is zero below $\omega = 1$. This corresponds to the estimate made earlier in the chapter that the input signal components in antiaircraft work lie roughly in the band below about 0.1 or 0.2 cycle per second. Let it be supposed also that A at higher frequencies is equal to 3 népers, corresponding to an average amplitude reduction of about 20

^c This is not intended to imply that the distinction between the initial waiting period and the "true transient" period is quite as sharp as it is in Figure 10. The selectivity in a data-smoothing filter is usually not great enough to justify the assumption that components beyond the linear phase region are of negligible importance.

to 1. Then $dB/d\omega$ at the origin is given from equation (3) as $6/\pi$ seconds, and in accordance with the rule just enunciated the minimum delay to be expected from such a structure in a data-smoothing application would consequently be $12/\pi$ seconds.

Aside from such specific quantitative relations equation (3) is useful as a basis for a number of important qualitative conclusions. One, for example, is the fact that although a lag is a necessary concomitant of any system showing a high-frequency attenuation, the amount of the lag depends greatly upon the portion of the frequency spectrum in which the attenuation is found. Since the integral is taken on an inverse frequency scale, a small attenuation at low frequencies is much more important than a considerably greater attenuation further out in the spectrum. This points to the desirability of designing tracking instruments which generate principally high-frequency noise, even if the amplitude of the noise is somewhat increased thereby. We may also notice that since the attenuation is a logarithmic function of amplitude an initial moderate reduction in the amplitude of disturbing noise may be much less expensive in lag than subsequent attempts at further reduction. For example, an amplitude reduction from 100 to 10 per cent over a given portion of the frequency spectrum produces no more lag than a subsequent reduction from 10 to 1 per cent.

5.6 WIENER'S PREDICTION THEORY— ZERO NOISE CASE

In Chapter 7 we distinguished between what we called the simple data-smoothing problem and the data-smoothing and prediction problem. The simple problem, with which this report is chiefly concerned, is the one which has been given principal attention thus far. On account of its broad interest, however, it seems worth while to include also a brief statement of Wiener's solution of the general problem. The method of development used here is intuitive and nonrigorous in comparison with Wiener's own development, but it permits the principal relations to be established by very elementary means.

It is convenient to consider first the zero noise case. The past history of the signal, then,

is known perfectly, and the existence of a prediction problem depends entirely upon the fact that since the signal is assumed to be statistical in character, its future is not completely determined from its past. The situation can be thought of in the terms suggested by Figure 11. The actual signal output appears at

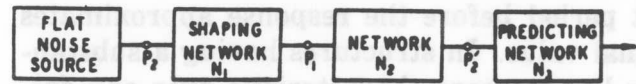


FIGURE 11. Schematic representation of Wiener's prediction theory when there is no noise.

P_1 . In accordance with the discussion earlier in the chapter, we imagine this signal to be generated by passing flat noise through the shaping network N_1 . The transfer admittance $Y_1(i\omega)$ of N_1 is determined from the power spectrum of the signal by the procedure outlined earlier and is a minimum phase shift characteristic. It will be recalled that minimum phase shift transfer admittances have the important property that their reciprocals are also the transfer admittances of physically realizable networks.

From Y_1 we can readily compute the transient response characteristic of N_1 . We shall assume for illustrative purposes that the impulsive admittance of N_1 takes the special shape shown by Figure 12.

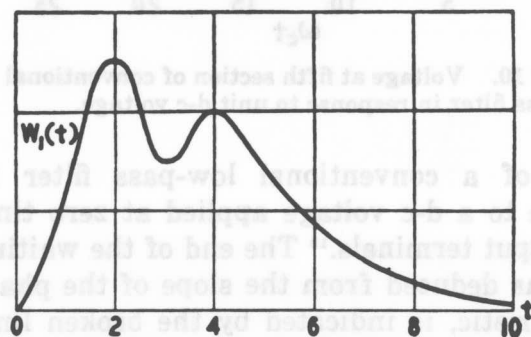


FIGURE 12. Assumed impulsive admittance of shaping filter.

The flat noise is thought of as consisting of a large number of elementary impulses with random amplitudes and occurring at random times. For the purposes of this analysis, however, it is sufficient to consider only the three unit impulses shown in Figure 13. Impulse B is supposed to occur at the instant at which

the prediction is to be made, A occurs two seconds in the past, and C , one second in the future. The response of N_1 to these three impulses will evidently be three curves of the sort given by Figure 12, suitably displaced in time as shown by Figure 14.

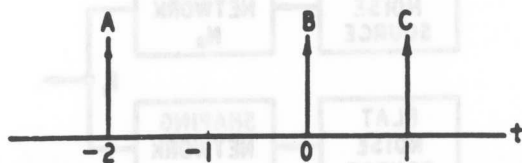


FIGURE 13. Impulses giving rise to applied signal through shaping filter.

The desired output of the predicting network is the curve of Figure 14 advanced by the prediction time, which we can assume, for illustration, to be two seconds. It may be assumed

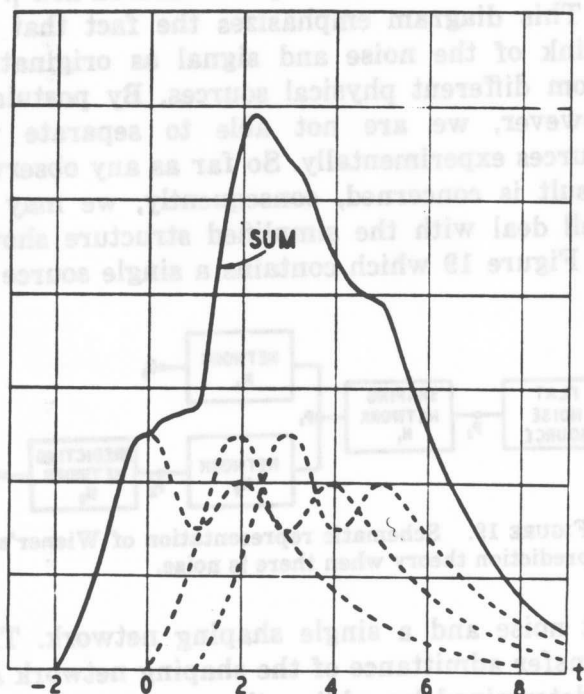


FIGURE 14. Applied signal at P_1 .

for the sake of preliminary analysis that the input of the predicting network is the three original impulses of Figure 13. The terminal P_2 at which they are supposed to appear is of course a purely fictitious one and is not accessible to us physically. We can, however, construct the equivalent terminal P'_2 by imposing the actual signal from terminal P_1 on the network N_2 , whose transfer admittance is the reciprocal of that of N_1 .

Let the predicting network connected to terminal P'_2 be represented by N_3 . Obviously a perfect prediction would be secured if N_3 could be assigned the impulsive admittance shown in Figure 15, that is, an impulsive admittance

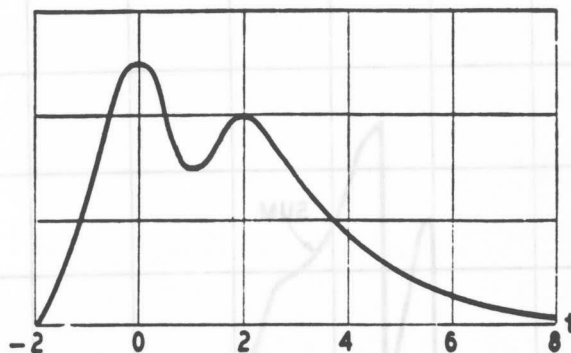


FIGURE 15. Ideal impulsive admittance of prediction network N_3 in Figure 11.

equal to the impulsive admittance of the original network but moved forward by the 2-second prediction time. Then all the constituent curves and the sum curve in Figure 14 would similarly be moved forward. Of course we cannot assign N_3 an impulsive admittance which is different from zero at negative times without postulating a nonphysical network. It is, however, perfectly possible to define N_3 from the portion of the impulsive admittance characteristic at positive times, with the remainder set equal to zero. This gives an impulsive admittance of the type shown by Figure 16. When energized by the three unitary impulses, it gives the result shown in Figure 17. The contributions of impulses A and B are not affected by the absence of a negative time portion of the impulsive admittance, but the contribution of impulse C is lost.

To formulate a physical prediction network

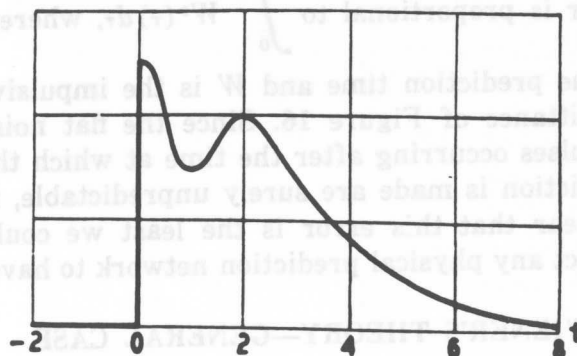


FIGURE 16. Realizable portion of required impulsive admittance.

we have merely to find by conventional methods the steady-state admittance Y_3 corresponding to the impulsive admittance of Figure 16. The two networks N_2 and N_3 may then be

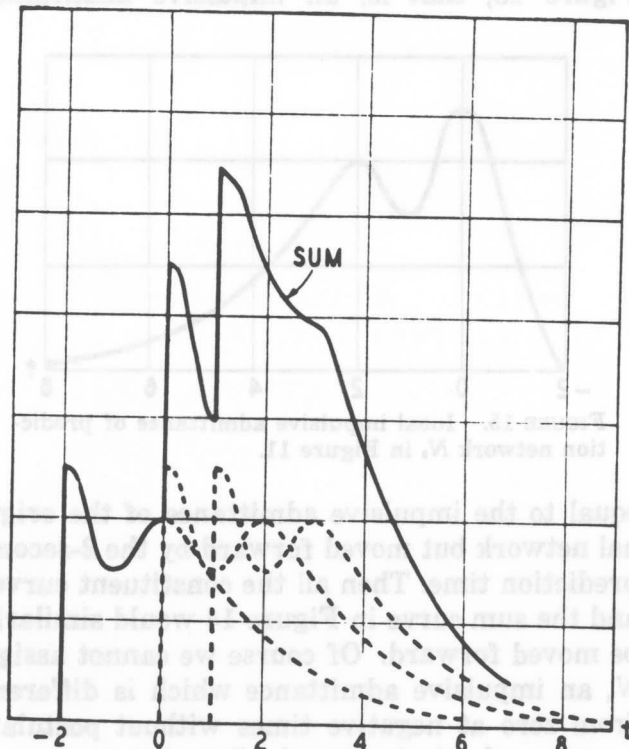


FIGURE 17. Response of realizable prediction network.

combined to give a single structure with the transfer admittance $Y_3 Y_2 = Y_3 / Y_1$, which will give the complete prediction when energized by the actual signal.

The mean square error in prediction is easily determined from the fact that the contributions of all impulses of the sort represented by C , occurring in the prediction interval, are lost. Since impulses in the flat noise source occur at random times the mean square

error is proportional to $\int_0^x W^2(\tau) d\tau$, where α

is the prediction time and W is the impulsive admittance of Figure 16. Since the flat noise impulses occurring after the time at which the prediction is made are surely unpredictable, it is clear that this error is the least we could expect any physical prediction network to have.

8.7 WIENER'S THEORY—GENERAL CASE

When the input data includes noise as well as the signal it is natural to think of the situation

in the manner shown by Figure 18. The first source of flat noise, together with the shaping network N_a , is the combination we have already used to represent the signal in the noise-free

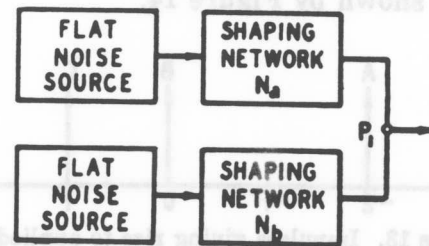


FIGURE 18. Circuit representation of random functions representing signal and noise.

case. The addition of noise is represented by the second independent source of flat noise with its associated shaping network N_b . They combine to give the total input measured at P_1 .

This diagram emphasizes the fact that we think of the noise and signal as originating from different physical sources. By postulate, however, we are not able to separate the sources experimentally. So far as any observed result is concerned, consequently, we may as well deal with the simplified structure shown in Figure 19 which contains a single source of

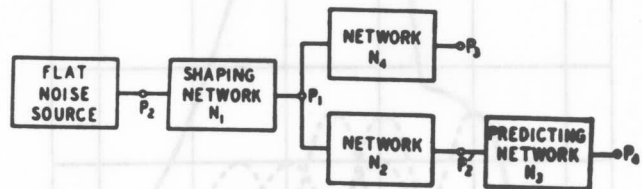


FIGURE 19. Schematic representation of Wiener's prediction theory when there is noise.

flat noise and a single shaping network. The transfer admittance of the shaping network N_1 is determined by adding the power spectra of signal and noise, converting the result to an amplitude characteristic, and computing the corresponding minimum phase according to the methods already used for the noise-free case.^d

Although we cannot separate the signal from

^d Note that the shaping network thus obtained is not the same as the one we would secure by adding the transfer admittances of N_a and N_b in Figure 18 directly. In order to realize the same total power at P_1 in each case, it is necessary to begin by adding the powers rather than the amplitude characteristics associated with the two paths.

the noise completely, we saw earlier that the mean square difference between the total input and the signal is minimized if we multiply the amplitude of the input at each frequency by the ratio of the signal power to the sum of the signal and noise powers. A fictitious filter having the prescribed amplitude characteristic is represented by N_1 in Figure 19. We assigned N_1 a zero phase characteristic so that there may be no lag in producing the result at P_3 . Thus the output at P_3 at any instant represents the best conceivable estimate (in the least squares sense) of the signal at that instant. The assumption of zero phase, of course, makes N_1 nonphysical, since it must have at least the minimum phase characteristic associated with its prescribed amplitude characteristic. This, however, is not an objection here since the structure is introduced purely for purposes of analysis.

The situation is now reduced to a form in which it is substantially equivalent to the one appearing in the zero-noise case. We assume a series of random impulses at P_2 which would produce responses at P_3 . The problem is that of advancing the response to each impulse so that the same result appears α seconds earlier at terminal P_1 . The solution is represented by networks N_2 and N_3 , which discharge functions similar to those of the correspondingly labeled networks in Figure 11. Thus, the network N_2 is the reciprocal of N_1 and is provided to make terminal P'_2 equivalent to P_2 as a source of impulses. Network N_3 is defined by an impulsive admittance obtained from the impulsive admittance between P_2 and P_3 by advancing the latter characteristic α units in time and then discarding the portion at negative time.

In this procedure there is only one point at which the situation differs from that without noise. In the noise-free case, the original impulsive admittance which we wished to advance in time was identically zero at negative times. In order to secure a physically realizable result, we needed only to discard the portion of the impulsive admittance between $t = 0$ and $t = \alpha$. In the present situation, on the other hand, the impulsive admittance is taken from a path including the nonphysical network N_1 . Thus the admittance may be expected to take such form as that shown in Figure 20, with nonzero am-

plitudes at both negative and positive times, and in order to secure a physical final network it is necessary to discard everything to the left of the line α .

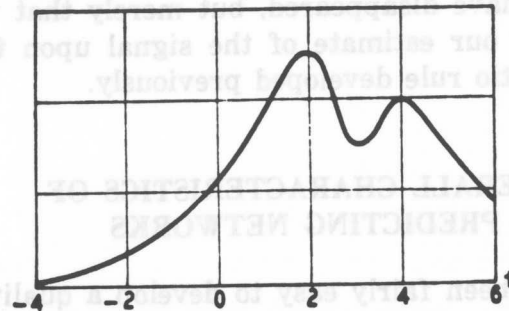


FIGURE 20. Typical impulsive admittance of best smoothing network N_1 in Figure 19.

This difference in the impulsive admittance characteristics has two consequences. The first is the fact that since the uncertainty of the prediction is measured by the amount of impulsive admittance which must be discarded, it is evidently greater in the present case where we are discarding much more. The second is the fact that in the noise-free case uncertainty exists only for a positive prediction time. A negative prediction time, which corresponds, of course, to the determination of the value assumed by the signal at some time in the past, can be set into the analysis as easily as a positive prediction time, merely by shifting the impulsive admittance to the right rather than the left. In the noise-free case, however, there is nothing to be discarded when we shift to the right, since the impulsive admittance with which we begin is in any case identically zero for negative times. Thus the uncertainty in the determination of any past value of the signal is zero. Since we have postulated no noise to confuse the data, this is, of course, an inevitable result. As soon as noise is included, on the other hand, there is no such sharp distinction between the future and the past.* The uncertainty in the determination of the true value of the signal in the near past is almost as great as it is in estimating what the signal will be in the near future. As we go further

* This statement is to be understood in a physical rather than a mathematical sense. It is not intended to imply that there may not be sharp changes of behavior in the impulsive admittance at zero.

and further into the past the uncertainty gradually diminishes. If we can allow ourselves unlimited lag, we at length reach a point at which the discarded portion of the impulsive admittance characteristic is negligibly small. This, however, does not mean that all uncertainties have disappeared, but merely that we can base our estimate of the signal upon the power-ratio rule developed previously.

8.8 OVERALL CHARACTERISTICS OF PREDICTING NETWORKS

It has been fairly easy to develop a qualitative picture of the general characteristics of typical data-smoothing networks. As we have seen, they have amplitude characteristics of the low-pass filter type combined with lagging phase shifts. No corresponding qualitative picture of the characteristics of a typical overall predicting circuit has, however, been developed as yet. The discussion just concluded provides a rule for determining the characteristics of a predicting circuit in any given case, but provides comparatively little in the nature of a description of the result we may expect to secure.

In any particular situation we can, of course, calculate the overall characteristics of the predicting circuit. A simpler way of characterizing the overall predictor characteristic qualitatively, however, is based upon the use of the attenuation-phase relations for physical networks. We need merely use such an equation as (3) backward. Thus, we have previously shown that a positive phase slope corresponds to a lagging output. Correspondingly, a negative phase slope can be interpreted to represent a lead, or in other words, a prediction.¹

¹ This, of course, does not mean that a network with a negative phase slope can predict a perfectly arbitrary event. We can hope to realize a negative phase slope, in combination with a flat amplitude characteristic, over only a finite band. The spectrum of an arbitrary event, that is, any suddenly applied signal, will always include important components running out to infinite frequency, where the negative phase slope can no longer be realized. The statement does, however, mean that if we suddenly apply a signal made up of one or more low-frequency sinusoids, and wait for the steady state to become established, the output will appear to lead the input by a time equal to the slope of the negative phase characteristic.

If we assign $(dB/d\omega)_{\omega=0}$ in equation (3) a negative value, we see that $A - A_0$ must on the average be negative. In other words, the amplitude characteristic of an overall prediction circuit must rise, on the average, as we proceed upward from zero frequency. This is in marked contrast to a data-smoothing network, which, as we have seen, tends to have a low-pass filter type of characteristic with a falling amplitude characteristic at high frequencies. The increased amplitude of response may have two detrimental effects. In the first place, it evidently produces a distorting effect on any signal components to which it applies. In the second place, it produces an exaggerated response to noise.

Examples of the characteristics of overall prediction circuits are readily constructed by reference to the circuit of Figure 21. Various

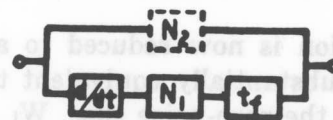


FIGURE 21. One-dimensional prediction circuit with data-smoothing networks.

particular results are obtained by assigning particular characteristics to the data-smoothing network. Thus, if the data-smoothing network is absent entirely the transmission through the path containing the differentiator is $i\omega t_1$, since differentiation is equivalent to multiplication by $i\omega$. The attenuation of the overall circuit is consequently $A = -\log |1 + i\omega t_1|$. This is plotted as curve I of Figure 22. The increasing amplitude characteristic at high frequencies is obviously due fundamentally to the increased transmission through the differentiator circuit.

If the data-smoothing network is assigned the characteristic $(1 + i\omega\alpha)^{-1}$, corresponding to a very simple low-pass filter type of response, the overall transmission becomes that shown by curve II in Figure 22. (It is assumed that $\alpha = t_1$, for simplicity.) The negative attenuation at high frequencies is much reduced. This is paid for by an increased amplitude of response at low frequencies, but since the integration in (3) takes place on an inverse frequency scale, the low-frequency fragment is much less than the gain reduction at high frequencies. Curve

III shows the result when the data-smoothing network is assigned the characteristic $(1 + i\omega a)^{-2}$. Finally, curve IV shows the result obtainable when there is also a filter in the

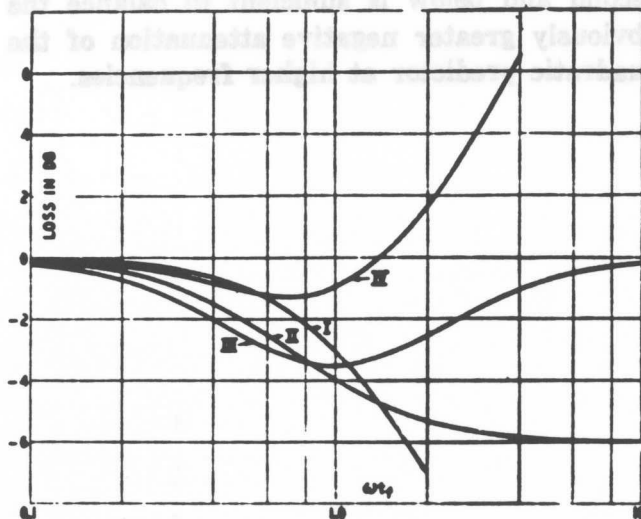


FIGURE 22. Attenuation characteristics of prediction circuit shown in Figure 21.

present-position circuit (as shown by the broken lines in Figure 21), so that there may be a net positive attenuation at high frequencies.

In view of the inverse frequency scale in (3), the gross negative attenuation will be minimized if the negative attenuation region is placed very close to zero frequency. This, however, means that much of the signal energy falls in the negative attenuation region so that in certain respects, at least, the signal response must be seriously injured. For example, in the specific circuits just discussed we can place the negative attenuation region at very low frequencies by choosing very long time constants, a , in the data-smoothing networks, with the consequence that the circuits will operate correctly for any long continued straight line path, but will be very sluggish in changing from one straight line to another. If the negative attenuation region is placed at higher frequencies, on the other hand, the signal response is improved but beyond certain limits the circuit becomes unbearably sensitive to noise.

Quantitative illustrations of these relationships are quickly constructed. Suppose, for example, that the prediction time is 2 seconds. From (3) this is consistent with an attenua-

tion characteristic having zero attenuation below $\omega = 1$ and a net gain of π népers thereafter. In other words, the amplitudes of all frequencies below $\omega = 1$ are increased by a factor of about 22 to 1. If the region of added gain is pushed to a higher frequency or concentrated within a narrow band, the multiplying factor rapidly becomes larger. For example, if we maintain A at approximately zero below $\omega = 2$, the average gain above this point must be 2π népers, corresponding to a multiplying factor of 500 to 1. We secure the same factor by attempting to concentrate the region of negative attenuation in the band between $\omega = 1$ and $\omega = 2$. The multiplying factor also goes up rapidly as we increase the prediction time. For example, with the gain uniformly spread over the frequency region above $\omega = 1$ the multiplying factor is 500 for a prediction time of 4 seconds, or more than 10,000 for a prediction time of 6 seconds.

Reasonable multiplying factors with long prediction times can be obtained only by carrying the negative attenuation region to very low frequencies. As indicated previously, the cost of this is an increase in the time required for the signal to change from one constant or nearly constant value to another. For example, in the first illustration above, if the region of π népers net gain is carried down from $\omega = 1$ to $\omega = 0.2$ the integral in (3) is just five times as great as it was before, so that the characteristic corresponds to a prediction time of 10 rather than 2 seconds. This change would correspond to an increase* from perhaps 4 or 5 to perhaps 20 or 25 seconds in the time required for the circuit to settle from one constant value to another.

Practical examples of the transmission characteristics of overall prediction circuits, with particular emphasis on the dominant effect of even very small negative attenuations at extremely low frequencies, are shown later in Figures 5 to 8, inclusive. In the linear predictor, $A - A_0$ varies as $-k\omega^2$ nears zero, and it is easily seen that such a term makes a finite con-

* Only rough numbers can be given, since circuits with the square-cornered attenuation characteristics chosen for illustrative purposes would have very ripply transient characteristics, corresponding to no very well marked settling time.

tribution to the integral in (3). On the other hand, the attenuation of the quadratic predictor, which is capable of dealing exactly with polynomial functions of time of the second degree or less, is necessarily zero at the origin^b

^b Cf the discussion of Quasi-Distortionless Prediction Networks in Appendix A.

to terms of the order of ω^4 , so that the integral in this region can be neglected. This slight difference between the two characteristics at frequencies of the order of 0.01 cycle per second and below is sufficient to balance the obviously greater negative attenuation of the quadratic predictor at higher frequencies.

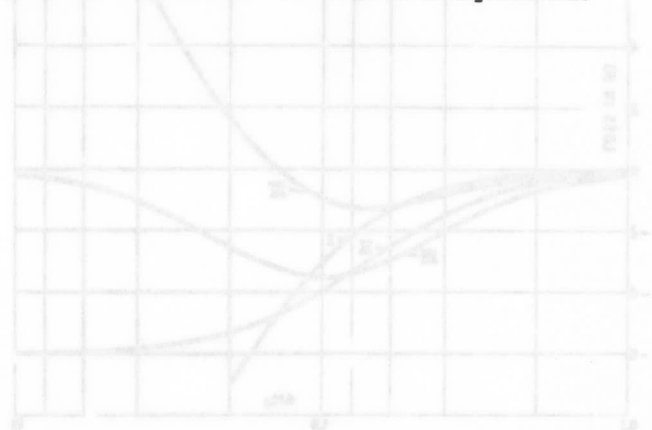


FIGURE 32. Attenuation characteristics of prediction circuit shown in Figure 31.

present-position circuit (as shown by the broken lines in Figure 31), so that there may be a net positive attenuation at high frequencies.

In view of the inverse frequency scale in (3), the gross negative attenuation will be minimized if the negative attenuation region is placed very close to zero frequency. This, however, means that much of the signal energy falls in the negative attenuation region so that in certain respects, at least, the signal response must be seriously injured. For example, in the specific circuits just discussed we can place the negative attenuation region at very low frequencies by choosing very long time constants, in the data-smoothing networks, with the consequence that the circuits will operate correctly for any long continued straight line path, but will be very sluggish in changing from one straight line to another. If the negative attenuation region is placed at higher frequencies, on the other hand, the signal response is improved but beyond certain limits the circuit becomes unduly sensitive to noise.

Quantitative illustrations of these relationships are quickly constructed. Suppose, for example, that the prediction time is 2 seconds. From (3) this is consistent with an attenu-

Practical examples of the transmission characteristics of overall prediction circuits, with particular emphasis on the dominant effect of even very small negative attenuations at extremely low frequencies, are shown later in Figures 5 to 8, inclusive. In the linear prediction, A varies as ω^2 , nearly zero, and it is easily seen that such a term makes a finite constant value to another.

Only rough numbers can be given, since circuits with the same overall attenuation characteristics chosen for illustrative purposes would have very different settling times.

THE ASSUMPTION OF ANALYTIC ARCS

THE DISCUSSION in the previous two chapters has been based upon the assumption that the least squares criterion forms a suitable measure of performance for a predicting network. This assumption permitted us to restrict our attention to the amplitude spectra of the signal and noise, leaving phase relations entirely out of account. Thus, both signal and noise could be thought of as "random noise" functions characterized by random phases and Gaussian distributions, as described in the preceding chapter. So far as the noise is concerned, there seems to be nothing wrong with this assumption. In the case of the signal, however, it appears that significant phase relations may exist. This chapter will consequently set up an alternative analysis which permits the significance of possible phase relations in the target paths to be estimated.

The alternative analysis is based upon the assumption that the target courses are sequences of analytic segments of different lengths joined together. These segments are simple predictable curves such as straight lines, parabolas, and circles. Significant phase relations are implied by the assumption that there are sudden changes from one type of course to another.

This picture of target paths is, of course, extreme. There are no such sharp discontinuities between one segment and another, nor do airplanes fly perfectly along simple curves even for limited periods. Nevertheless, it is the conception of target courses upon which the rest of our analysis is based. The reasons for believing that it is a closer approximation to actual target courses than, say, a random noise function with the same power spectrum would be, are given later. Perhaps more important is the fact that the possibility of hitting an airplane flying along such a simple analytic arc is much greater than it would be if we were attempting to predict a corresponding random noise function. It is thus advantageous to take the analytic arc assumption as a basis for designing the prediction circuit,

even if the assumption seems to be reasonably well justified over only occasional segments of actual target paths. An example of such a situation is furnished by the bombing run illustration described in Chapter 7.

As a corollary to the analytic arc assumption it is also assumed that the theoretical predicted point must be quite close to the actual target position if the probability of scoring a hit is to be appreciable. In other words, such dispersive factors as random errors in computer or gun or the lethal radius of the shell, which would tend to produce occasional hits at long distances from the theoretical predicted point, are quite small. This is such a plausible assumption in the light of present-day antiaircraft experience that its critical importance in the present argument is likely to go unperceived. However, this is the assumption which limits consideration to small errors in prediction, whereas the least squares criterion naturally gives greatest emphasis to large errors. If, for example, antiaircraft projectiles were suddenly endowed with a much greater destructive radius, we would be much more interested in fairly large misses, and the objections to the least squares criterion would disappear.

These postulates are discussed in more detail in the following sections. In anticipation of this discussion the following conclusions may be mentioned:

1. With the assumptions as stated, the prediction should be on a modal rather than a least squares basis. In other words, the gun should be aimed at the most probable future position of the target.

2. Modal prediction requires evaluation of the parameters of the analytic arc the target is at present traversing. This can be accomplished by smoothing the values of these parameters evaluated for a period in the past.

3. If the smoothing is performed by linear invariable networks, the impulsive admittances of these networks should have a definite cutoff after a finite smoothing time. By this means

all data over a certain age are given zero weight. The method of calculating the proper smoothing time is developed.

4. Definite advantages can be obtained from circuits with variable smoothing times if such systems can be satisfactorily mechanized.

9.1 THE TARGET COURSES

The target courses, like the tracking errors, can be thought of as a statistically generated set of functions—that is, a stochastic process. The structure of this process is, however, very different from that of the tracking errors. It is by no means satisfactory to assume the target courses to be equivalent to a random noise having the same power spectrum as the target courses. As we pointed out in Chapter 7, the target is piloted by a purposeful human being. It tends to follow a definite simple curve for a period of time and then to shift to a new simple curve. Much of the flight is in attempted straight lines with constant velocity. Most of the remainder can be considered to be segments of circles or helices in space, or as segments of parabolas or higher degree curves. Straight line constant speed flight corresponds to the airplane controls in a neutral position. The helical flight is a natural generalization allowing arbitrary, but fixed, positions of the controls. The curves which are parabolic functions of time correspond to constant acceleration in the three space coordinates. Thus, all these assumptions have a reasonable physical background.

Most antiaircraft computers are constructed on the assumption of straight line flight, although some work has been done in World War II on curved flight directors both with the helical and the parabolic assumptions. There is not a great deal of difference in these two generalizations from the practical point of view, since determination of acceleration terms is subject to such large errors in any case.

The important part of this representation of the target courses is that they consist of segments of simple analytic curves joined together. The individual segments are completely predictable if we have a part of the segment given exactly. One need merely evaluate the parameters of the segment from the given part

and evaluate the curve for $t = t_f$. The unpredictable part of the target courses is due to the possibility of sudden changes from one segment to another. With random noise functions the unpredictableness occurs continuously.

This simplified description of the target courses as piecewise analytic functions must be recognized as only a first approximation. A more complete description of the target course would include the "fine structure," the connecting curves between the various analytic segments and the deviations from the segments due to random air disturbances and similar causes. This latter effect, the wandering of the target from its intended path, might be reasonably well represented by the addition of a random noise function to the piecewise analytic functions described above.

9.2 THE POISSON DISTRIBUTION OF SEGMENT END POINTS

The analytic segments of which the course is supposed to consist are not all of the same duration—we may assume some probability distribution of the duration of these segments. The simplest assumption here is that the breaks occur in a Poisson distribution in time. This assumption is not necessary for our analysis but is a reasonable one and leads to a simple mathematical treatment. Any other reasonable distribution would give comparable results.

A series of events is said to occur in a Poisson distribution in time if the periods between successive events are independent in the probability sense and are controlled by a distribution function

$$p(l)dl = \frac{1}{a} e^{-l/a} dl.$$

Here $p(l)dl$ is the probability of an interval of length between l and $l + dl$. This means that the frequency of intervals of a given length is a decreasing exponential function of the length. This type of distribution is familiar in physics as describing the decay of radioactive substances. The time a in the distribution function is the average length of the intervals, since

$$\bar{l} = \int_0^{\infty} l p(l) dl$$

$$= \int_0^{\infty} \frac{l}{a} e^{-l/a} dl$$

$$= a.$$

It is related to the "half life" b of the interval by

$$b = a \ln 2.$$

The single number a completely specifies the Poisson distribution. The events may be said to be happening as randomly as possible apart from the fact that they occur at an average rate of $1/a$ per second.

Another way of describing a Poisson distribution of events is the following. The probability of an event in a small interval of duration dl is $(1/a)dl$ and is independent of whether or not events have occurred in any other nonoverlapping intervals.

9.3 THE PROBABILITY DISTRIBUTION OF FUTURE POSITIONS

Let us suppose that we have a record of the course of the target up to the present time and a complete statistical description of the set of target courses. What can then be said about the position of the target t_i seconds from now? If we were able to analyze the data completely the most we could obtain would be a probability distribution function for the future position. This distribution function would give the probability, in the light of the course history, of the target being at any point in space at the future time. This function would assume large values at likely points and low values at unlikely points. For t_i small the distribution would be highly concentrated and for larger t_i it would tend to spread out.

In the simple case we have been discussing, of a Poisson distribution of sudden changes in type of course, the distribution consists of two parts. First, there is a spike of probability at one point, the continuation of the present predictable segment. Second, there is a continuous distribution which corresponds to possible changes to a new segment during the time of flight. As t_i increases the total probability in the spike decreases exponentially toward zero, and the total in the continuous part increases exponentially toward unity. The behavior is roughly as indicated in Figure 1.

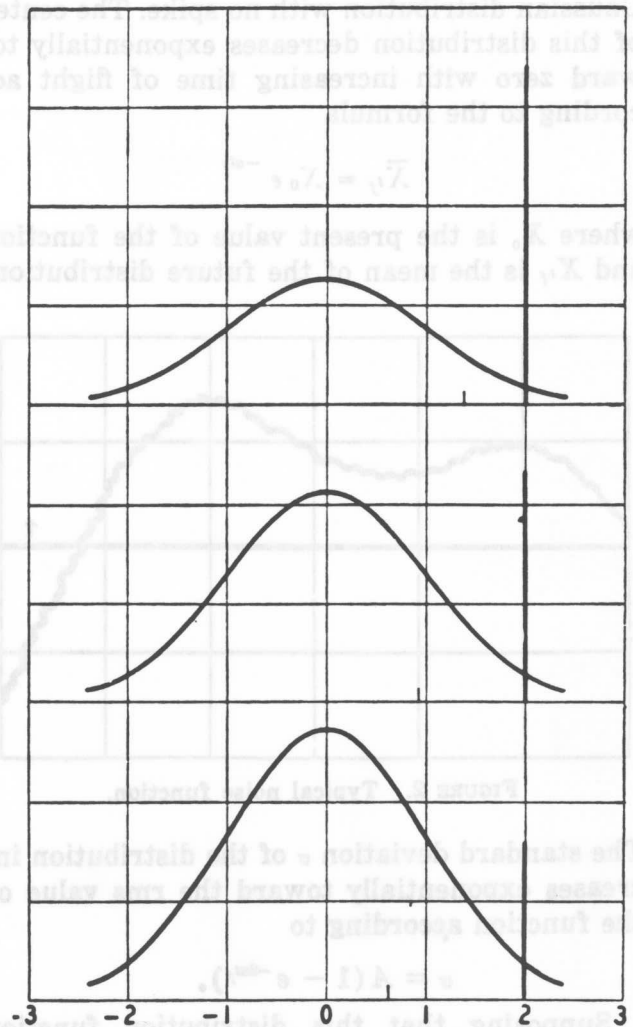


FIGURE 1. Probability distribution of future position of target, assuming piecewise analytic courses.

A very different type of future position distribution is exhibited with other assumptions about the target courses. For example, suppose the courses were random noise functions with the power spectrum

$$P(\omega) = \frac{1}{a^2 + \omega^2}.$$

A typical noise function with this spectrum is shown in Figure 2. In Figure 3 is shown a typical velocity under the other assumption, that the courses are piecewise analytic and in fact straight lines between breaks. If the breaks are Poisson distributed, both Figure 2 and Figure 3 have the same power spectrum, $1/(a^2 + \omega^2)$. The future distribution of velocities for Figure 3 is shown in Figure 1, and for Figure 2, it will be as shown in Figure 4. In the random noise case the future distribution is a

Gaussian distribution with no spike. The center of this distribution decreases exponentially toward zero with increasing time of flight according to the formula

$$\bar{X}_{t_f} = X_0 e^{-at_f}$$

where X_0 is the present value of the function and X_{t_f} is the mean of the future distribution.

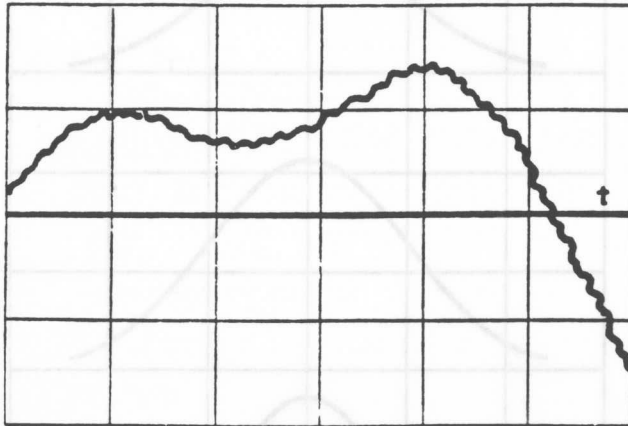


FIGURE 2. Typical noise function.

The standard deviation σ of the distribution increases exponentially toward the rms value of the function according to

$$\sigma = A(1 - e^{-2at_f}).$$

Supposing that this distribution function could be determined, where should the gun be aimed? The answer to this will depend on two factors: the gun dispersion, and the lethal

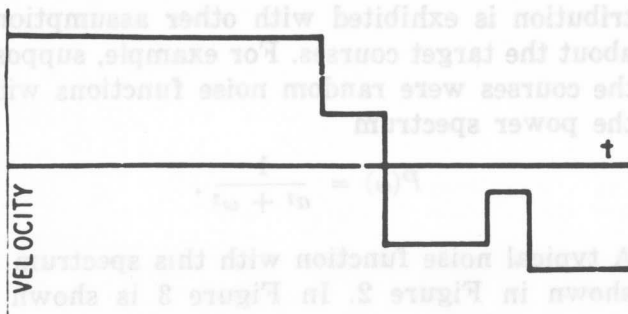


FIGURE 3. Typical velocity function.

effects of the shell. If the gun is aimed to explode the shell at a certain point in space, the shell will not necessarily explode at that point, but rather there will be a distribution of positions centered about the point aimed at, because of gun dispersion. Also, if the shell explodes at a certain point and the target is at

another point, there will be a certain probability of lethal effect which decreases rapidly with increasing distance between the points. These two functions could be combined by a product integration to give the probability of lethal effect if the target is at one point and

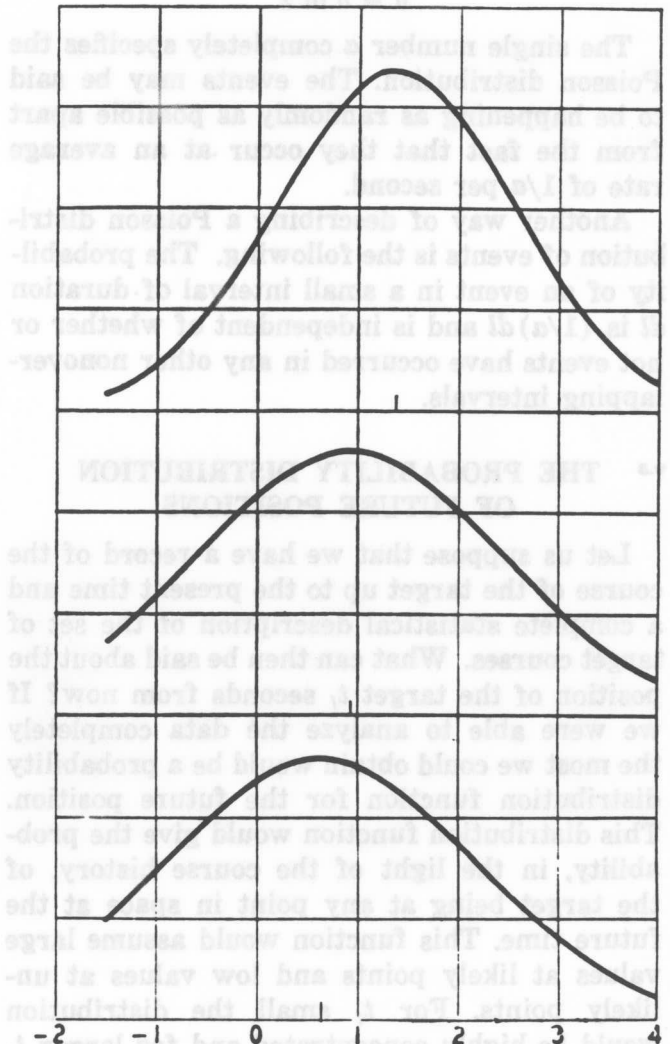


FIGURE 4. Probability distribution of future position of target, assuming courses with random noise properties.

the gun aimed to explode the shell at a second point. To determine the probability of a hit when aiming at a certain point, then, we should multiply the probability of the target being at each point in space by the probability of lethal effect when it is at that point and integrate the product over all space. The optimum point of aim will be the one which maximizes this integrated product.

In one dimension this may be expressed mathematically as follows. Let $P(x)$ be the

future position distribution of the target, so that $P(x)dx$ is the probability of it being in the interval from x to $x + dx$ at the future time. Let $Q(x,y)$ be the probability of hitting the target if the gun is aimed at point y and the target is at point x . Then the total probability of a hit when aiming at point y is

$$R(y) = \int P(x) Q(x,y) dx.$$

The point of aim y should be chosen to maximize $R(y)$.

In the cases we consider, the lethal radius of the shell and the dispersion of the gun are both assumed to be small in comparison with the range of future positions if there is a change of course during the time of flight. This means that $Q(x,y)$ is small unless x is very near to y . $Q(x,y)$ can be, in fact, considered to be a δ function of $(x-y)$, and the value $R(y)$ is then just a constant times $P(y)$. Thus, *the best aiming point under this assumption is the most probable future position of the target*. The assumption of small lethal distance is generally valid with antiaircraft fire and ordinary chemical explosive shells.

Now the most probable future position in our case is the spike of probability corresponding to the analytic extrapolation of the present segment of the target course. To determine its position one must find the parameters of this segment and evaluate for t_f seconds in the future. For example, if the segments are assumed to be straight lines (constant velocity target) the velocity components are determined and multiplied by t_f to give the predicted change in position. These changes are added to the present position to give the future position. If helical or parabolic segments are assumed, the parameters of these curves are determined from the past data, and the curves extrapolated t_f seconds into the future.

These conclusions may be contrasted with the idea of aiming at the point which minimizes the mean square error. The least squares criterion amounts to aiming at the mean or center of gravity of the future distribution of position. This point will ordinarily be under the continuous part of the distribution and not at the spike; e.g., the point marked in Figure 1. Its position depends to a considerable extent on

distant parts of the distribution, which would surely be complete misses in any case. The chief advantage of the least squares criterion is that it fits in well with the mathematical tools suitable to these problems, leading to solvable equations.

The least squares criterion will still appear in our analysis in that we attempt to smooth our *course parameters* in such a way as to minimize the mean square error in *these*, a very different thing from minimizing the mean square error in the predicted position of the target.

9.4 NECESSITY OF A SHARP CUTOFF

The changes in the course parameters between adjacent segments can be very large. Also, at the start of operations and in changing from one target to another there will be large and erratic variation of the input to the smoothing and predicting circuits, unrelated to the present target course. If any of these data are used in prediction, the result will almost surely be a miss because of the small lethal radius of the shell. The only way to eliminate these errors in a linear invariable system is to have all weighting functions cut off sharply after a short time. Then all data over a certain age are eliminated. Hits will occur only when the target has been on a predictable segment for this length of time or more and remains there at least t_f seconds in the future.

Suppose the weighting function for velocity has a 1 per cent tail beyond the cutoff point and that the trackers start following the target from a zero position. Then after the smoothing time there will be, because of the lack of exact cutoff, a 1 per cent error in velocity. If the time of flight were 15 seconds and the target velocity 200 yards per second, this represents an error of 30 yards in predicted position. Since this is comparable to the other errors in a typical director, we conclude that the tail of the smoothing curve should not be much greater than 1 per cent of its total area.

9.5 CALCULATION OF THE BEST SMOOTHING TIME

Under the assumptions we have made, the proper smoothing time to maximize the number of hits can be determined as follows. Let $P(l)$

be the probability that a predictable segment of the course lasts for l seconds or more. In the Poisson case this function is

$$P(l) = e^{-l/a}$$

With a given smoothing time S there will be a certain probability of hitting the target, assuming it has been on the present segment for S seconds in the past and will remain there for t_f seconds in the future. We assume changes in course to be so large that any change results in a miss. This probability of a hit $Q(S)$, provided it remains on the course, will be an increasing function of S . Ordinarily the standard deviation will decrease as the square root of the smoothing time. We have assumed the lethal radius of the shell small compared to the dispersion of shells about the target. The probability of a hit will then vary inversely with the volume through which the shells are dispersed. If the gun itself had no dispersion but all errors were due to tracking errors (and if the tracking error spectrum is flat), the probability of a hit would then vary as $KS^{3/2}$ for S in the region of interest. This is because there are three dimensions and the expected error in each of these is decreasing as $S^{-1/2}$. With gun dispersion present, $Q(S)$ will have the form

$$Q(S) = K \left(\sigma_1^2 + \sigma_2^2 \frac{a}{S} \right)^{-3/2}$$

where σ_1 is the standard deviation due to the gun dispersion, and $\sigma_2 \sqrt{a/S}$ that due to tracking errors. The sum of the squares is the total variance in each dimension and the three-halves power gives the total dispersion volume.

When these two functions $P(l)$ and $Q(S)$ are known, the best smoothing time is that which minimizes the product

$$P(S + t_f) \cdot Q(S).$$

The first term is the probability of a predictable segment of the course lasting $S + t_f$ seconds, and the second term is the probability of a hit if it does last that long. Therefore, the product is the probability of a hit with smoothing time S .

In the Poisson case, with no gun dispersion, the calculation is as follows:

$$P(l) = e^{-l/a}$$

$$P(S + t_f) = e^{-\frac{S+t_f}{a}} = Ae^{-S/a}$$

$$Q(S) = \sigma S^{3/2}$$

$$f(S) = P(S + t_f)Q(S) = Be^{-S/a} S^{3/2}$$

$$f'(S) = B \left[e^{-S/a} \frac{3}{2} S^{1/2} - \frac{1}{a} e^{-S/a} S^{3/2} \right] = 0$$

$$S = \frac{3}{2} a$$

The proper smoothing time is $\frac{3}{2}$ of the average segment length, and is independent of the time of flight and all other factors.

The presence of gun dispersion and computer errors which are independent of smoothing time decreases the best S from this value. In this case the equation for optimal S is the quadratic

$$\sigma_1^2 \left(\frac{S}{a} \right)^2 + \sigma_2^2 \frac{S}{a} - \frac{3}{2} \sigma_2^2 = 0;$$

hence

$$\begin{aligned} \frac{S}{a} &= \frac{-\sigma_2^2 + \sigma_2 \sqrt{\sigma_2^2 + 6 \sigma_1^2}}{2 \sigma_1^2} \\ &= \frac{\sigma_2}{2 \sigma_1} \sqrt{\left(\frac{\sigma_2}{\sigma_1} \right)^2 + 6} - \frac{1}{2} \left(\frac{\sigma_2}{\sigma_1} \right)^2 \end{aligned}$$

Here σ_1 is the part of the errors which is independent of smoothing time (dispersion errors in the computer, etc.) and σ_2 is the error which varies inversely with the square root of S , σ_1 being its value at $S = a$. Ordinarily σ_1 is several times σ_2 in which case we have approximately

$$\frac{S}{a} = \frac{\sigma_2}{\sigma_1} \sqrt{\frac{3}{2}}.$$

There are other factors which we have neglected, which decrease the best smoothing time still further. The wandering of the target about the predictable segments assumed in the above simplified analysis makes old data less reliable and therefore reduces S . Also, there is the tactical consideration that when starting to track a target it is desirable to commence firing as soon as possible, even if reducing this time makes individual hits somewhat less probable. For these and other reasons the best smoothing time will be just a fraction of a .

9.6 NONLINEAR AND VARIABLE SYSTEMS

The compromise required in choosing a certain definite smoothing time can be eliminated by the use of nonlinear elements. In particular, if a method is devised for determining when changes of course occur, this indication can be used to start a new linear but variable smoothing operation, so that the device uses all the data pertinent to the present segment and no data from previous segments. There is a clear improvement in such cases although not so great as might be expected. There are many practical difficulties in proper adjustment of such a "trigger" action. If the trigger is too sensitive it will assume new segments due merely to tracking noise and seldom allow sufficient smoothing for accurate fire. If it is too insensitive it fails in its function of quickly

locating changes of segment. Since the noise and target courses are subject to considerable variation, this adjustment is not easy.

In such a system the smoothing may be linear—the only nonlinearity is the tripping circuit. The analysis of best weighting functions, etc., given in later chapters can for the most part be applied to such cases. There may also be advantages to be derived from making the smoothing operator depend on the general position in space of the target relative to the gun. The smoothing time may be varied, for example, as a function of the time of flight. This type of variation would be slow compared to the noise frequency, and here again the linear analysis can be used.

Whether any real advantage can be obtained by "strongly" nonlinear smoothing in practical cases other than these two possibilities is questionable.

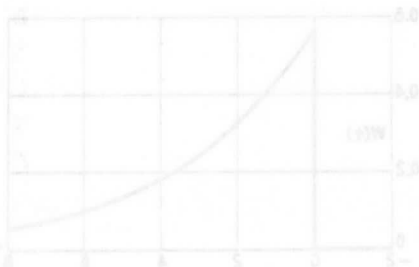


FIGURE 1. Simple exponential weighting function.

An impulsive admittance of the type shown in Figure 1 does not show any very definite settling time. The exponential curve approaches zero gradually, and it is a long time after a change in course before the effects of the data obtained on the old course are negligible. This is obviously an undesirable result.

* In exceptional circumstances the physical apparatus in which these processes are carried out may also be sources of additional noise.

With the settling time limit given, the problem of choosing a suitable data-smoothing network reduces to that of finding the best shape of the impulsive admittance characteristic for $t < T$. Obviously this shape determines how the output of the network changes in going from the parameter value appropriate for the first arc to that appropriate for the second. The exact way in which the response settles from one constant value to the next is, however, usually of comparatively little interest. The shape of the weighting function is of importance chiefly because of its effect on the noise. For each noise spectrum there is, in principle, an optimum shape for the weighting function. The present chapter approaches the problem of choosing a shape which will minimize the effect of noise from several points of view.

The data-smoothing network is most conveniently specified by its impulsive admittance. (See Appendix A.) In accordance with the assumptions made in the previous chapter, it will be assumed that the desired impulsive admittance is identically zero after some limiting time T . Thus T seconds after a change from one analytic arc to the next the new parameter value is established. T is the so-called "settling time" of the data-smoothing network.

SMOOTHING FUNCTIONS FOR CONSTANTS

THE ANALYTIC ARC ASSUMPTION described in the previous chapter immediately allows us to reduce a vast proportion of data-smoothing problems to a relatively concrete form. Obviously the arc will be specified by a number of parameters and the principal object of the computing and data-smoothing circuits must be to isolate values of these parameters on the basis of which a prediction can be made. In practical cases the instantaneous values of the parameters are isolated by coordinate converters. The function of the data-smoothing circuit is to provide a suitable average from these instantaneous values. This is called "smoothing a constant" here since the parameters are assumed to be constant along each arc, although they may change radically from one arc to another.

The data-smoothing network is most conveniently specified by its impulsive admittance. (See Appendix A.) In accordance with the assumptions made in the previous chapter, it will be assumed that the desired impulsive admittance is identically zero after some limiting time T . Thus, T seconds after a change from one analytic arc to the next the new parameter value is established. T is the so-called "settling time" of the data-smoothing network.

With the settling time limit given, the problem of choosing a suitable data-smoothing network reduces to that of finding the best shape of the impulsive admittance characteristic for $t < T$. Obviously this shape determines how the output of the network changes in going from the parameter value appropriate for the first arc to that appropriate for the second. The exact way in which the response settles from one constant value to the next is, however, usually of comparatively little interest. The shape of the weighting function is of importance chiefly because of its effect on the noise. For each noise spectrum there is, in principle, an optimum shape for the weighting function. The present chapter approaches the problem of choosing a shape which will minimize the effect of noise from several points of view.

It should be noted that the term noise as used here does not necessarily refer to the errors associated directly with the tracking data. The tracking data may have been subjected to coordinate conversions, differentiations, or other processes of computation before reaching the data-smoothing network.^a The noise associated with the signal to be smoothed thus will usually have characteristics differing from those of the noise associated with the tracking data.

10.1 EXPONENTIAL SMOOTHING

Before attacking the problem of smoothing a constant in a systematic way it is worth while to consider an important special case. This is the so-called exponential smoothing circuit. It leads to a data-smoothing network in which the output V is related to the input E by

$$V(t) = \alpha \int_0^{\infty} E(t - \tau) e^{-\alpha\tau} d\tau$$

so that the impulsive admittance $W(t)$ is an exponential function of time, as illustrated by Figure 1.

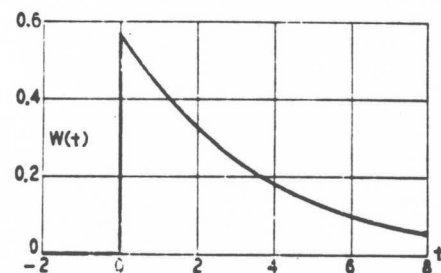


FIGURE 1. Simple exponential weighting function.

An impulsive admittance of the type shown in Figure 1 does not show any very definite settling time. The exponential curve approaches zero gradually, and it is a long time after a change in course before the effects of the data obtained on the old course are negligible. This is obviously an undesirable result,

^a In exceptional circumstances the physical apparatus in which these processes are carried out may also be sources of additional noise.

and the exponential weighting function is consequently not a recommended one for situations to which the analytic arc assumption applies. The exponential solution is, however, described here because it occurs in such a vast variety of cases. It is found, in fact, whenever the data-smoothing device is specified by a linear first-order differential equation with constant coefficients. It may thus correspond to many simple situations. For example, this is the result which would be obtained in an electrical circuit if we smoothed the data by placing a simple shunt capacity across a resistance circuit. In mechanical structures it is encountered whenever the damping depends either upon simple inertia or a simple compliance.

Simple exponential smoothing also occurs in a variety of other situations which may be somewhat less obvious. For example, it is the effective result in either an aided laying or a regenerative tracking scheme whenever the ratio between rate and displacement corrections is fixed. Another somewhat similar example is furnished by the feedback amplifier circuit shown in Figure 2. Since rapid fluctua-

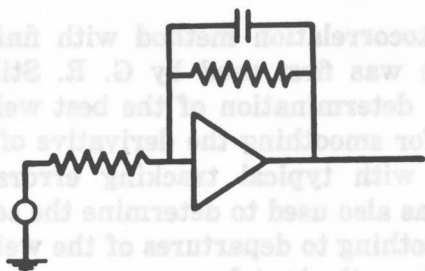


FIGURE 2. Feedback amplifier circuit giving simple exponential weighting function.

tions in the output of this amplifier are fed back through the capacity and tend to oppose the input voltage, the structure acts as a smoother, and more detailed analysis would show that it has characteristics similar to those obtained by using a shunt capacity across a resistance circuit. The structure is introduced here because considerable use is made of it in connection with the discussion of nonlinear smoothing in a later chapter.

One simple conclusion about data-smoothing networks can be drawn immediately from this discussion. Since all structures simple enough to be specified by a first-order differential equa-

tion give exponential smoothing, which has no very well-marked settling time, it is clear that a data-smoothing network which shows a well-defined settling time must probably be at least moderately complicated.

10.2

CURVE-FITTING METHOD

Consider the signal E shown in Figure 3 under the assumption that the true signal is constant and the superposed noise is random

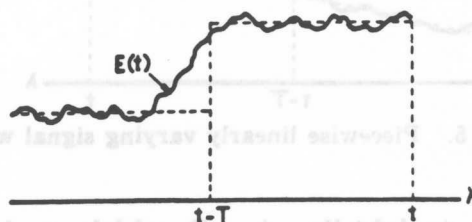


FIGURE 3. Piecewise constant signal with noise.

with a flat spectrum. The best constant A , in the least squares sense, which can be fitted to the signal from $t - T$ to t is that which minimizes

$$\int_{t-T}^t [A - E(\lambda)]^2 d\lambda,$$

viz.,

$$A = \frac{1}{T} \int_{t-T}^t E(\lambda) d\lambda. \quad (1)$$

Comparing this with equation (2), Appendix A, it will be seen that A , which is obviously a function of t , is the response to the assumed signal of a network whose impulsive admittance is

$$W(t) = \frac{1}{T} \quad 0 < t < T. \quad (2)$$

This is the best weighting function for smoothing under the assumed circumstances. It is illustrated in Figure 4.

A more complex situation is one in which the true signal is a line of constant slope with

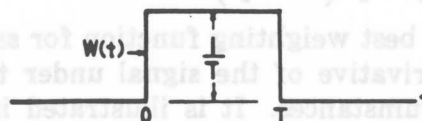


FIGURE 4. Best weighting function for smoothing piecewise constant signal.

superposed flat random noise, as shown in Figure 5. For convenience the analysis will be conducted in terms of the age variable $\tau = t - \lambda$.

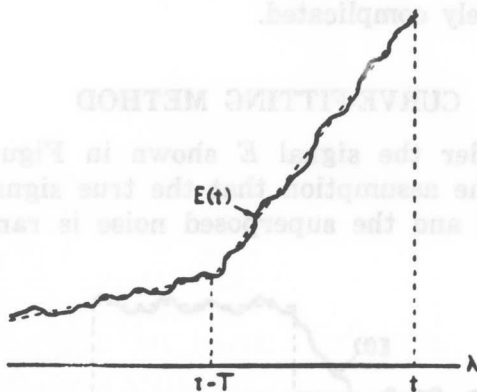


FIGURE 5. Piecewise linearly varying signal with noise.

The best straight line $A - B\tau$ which can be fitted to the signal from $\tau = 0$ to $\tau = T$ is that which minimizes

$$\int_0^T [A - B\tau - E(t - \tau)]^2 d\tau.$$

Hence A and B must satisfy simultaneously

$$\begin{aligned} A - \frac{T}{2} B &= \frac{1}{T} \int_0^T E(t - \tau) d\tau \\ \frac{T}{2} A - \frac{T^2}{3} B &= \frac{1}{T} \int_0^T E(t - \tau) \tau d\tau. \end{aligned} \quad (3)$$

Eliminating A , we get

$$B = \frac{12}{T^3} \int_0^T E(t - \tau) \left(\frac{T}{2} - \tau \right) d\tau,$$

whence by partial integration

$$B = \frac{6}{T^2} \int_0^T E'(t - \tau) \cdot \tau(T - \tau) d\tau.$$

Comparing this with (7), Appendix A, it will be seen that B , which is obviously a function of t , is the response to the derivative of the assumed signal of a network whose impulsive admittance is

$$W(t) = \frac{6}{T} \cdot \frac{t}{T} \left(1 - \frac{t}{T} \right) \quad 0 < t < T. \quad (4)$$

This is the best weighting function for smoothing the derivative of the signal under the assumed circumstances. It is illustrated in Figure 6 and is generally referred to as the "parabolic weighting function."

It should be noted also that the right-hand member of the first of equations (3) is formally the same as that of equation (1). Hence the response of the network specified by (2)

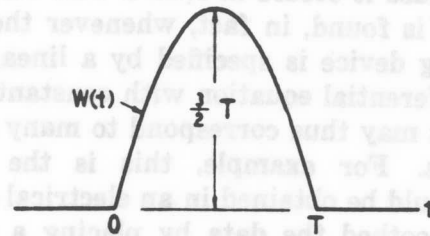


FIGURE 6. Best weighting function for smoothing piecewise linearly varying signal.

and illustrated in Figure 4, to the type of signal shown in Figure 5, will correspond to the value on the best straight line $T/2$ seconds back from t , the present time. This network is still the best for smoothing the signal, but it introduces a delay of one half of the smoothing time. The delay may be reduced only at the price of a reduction in smoothing unless the smoothing time is increased.

10.3 AUTOCORRELATION METHOD

The autocorrelation method with finite settling time was first used by G. R. Stibitz in numerical determination of the best weighting function for smoothing the derivative of tracking data with typical tracking errors. This method was also used to determine the sensitivity of smoothing to departures of the weighting function from the best form.

The analysis is based upon the formula

$$V(t) = \int_0^T g'(t - \tau) W(\tau) d\tau \quad t > T$$

for the response to the derivative of the error time function $g(t)$ of a network whose impulsive admittance or weighting function $W(t)$ is identically zero for $t > T$ as well as for $t < 0$. Since measured tracking errors are generally tabulated only at 1-second intervals, the integral may be approximated by the sum

$$V(t) = \sum_{m=1}^T \Delta g_{t-m+(\frac{1}{2})} W_{m-(\frac{1}{2})}$$

for integral values of t .

The instantaneous transmitted power is the

square of this expression, and the average transmitted power is

$$P_{\text{avg}} = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{t=0}^N V^2(t)$$

This may be expressed in the form

$$P_{\text{avg}} = \sum_{m=-1}^T \sum_{n=1}^T W_{m-(1/2)} \cdot C_{m-n} \cdot W_{n-(1/2)} \quad (5)$$

where

$$C_{m-n} = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{t=0}^N \Delta g_{t-m+(1/2)} \cdot \Delta g_{t-n+(1/2)}$$

is the autocorrelation of the errors. Having computed the autocorrelation, (5) may be minimized with respect to the W 's by familiar methods, under the constraint

$$\sum_{m=-1}^T W_{m-(1/2)} = 1.$$

The values of W thus obtained are the specification of the best weighting function.^b Equation (5) may then be used to determine the sensitivity of smoothing to departures of the weighting function from the best form.

Proceeding along this line, Stibitz found that the best weighting function for typical actual tracking errors was generally intermediate to the uniform and parabolic ones shown in Figures 4 and 6. Furthermore, Stibitz found that the difference in smoothing obtained from the best weighting function on the one hand and from the uniform or the parabolic weighting function on the other hand, is negligible in practice.

The autocorrelation method was later formalized by R. S. Phillips and P. R. Weiss who incorporated it into a theory of prediction.⁷ A brief exposition of this formulation is given in Appendix B.

10.4 ELEMENTARY PULSE METHOD

For the purposes of this method, an elementary noise pulse is defined by a time function $F_0(t)$ which satisfies the following requirements:

1. Identically zero when $t < 0$.

^b The computations involved may be considerably reduced by noting the symmetry property proved in Section B.2, Appendix B.

2. Contains no terms which increase exponentially with time.

3. Power spectrum $N(\omega^2)$ is the same as that of the noise.

The noise is then regarded as the result of elementary noise pulses started at random. Alternatively, it may be regarded as the result of flat random noise passed through a network whose transmission function is $S(p) = L[F_0(t)]$. As a matter of fact, only $S(p)$ is required in the analysis, and this is readily determined from the relation

$$|S(i\omega)|^2 = N(\omega^2),$$

together with the condition that $S(i\omega)$ corresponds to the transmission function of a minimum-phase physical structure (cf. Appendix B).

The response $F(t)$ to the elementary noise pulse $F_0(t)$ of a network whose impulsive admittance is $W(t)$ is given by the operational equation

$$F(t) = S(p) \cdot W(t)$$

in accordance with the footnote in Section A.5, Appendix A. The best form for $W(t)$ is therefore that which minimizes the integral

$$\int_{0-}^{\infty} [F(t)]^2 dt \quad (6)$$

under the restriction

$$\int_{0-}^{t_0} W(t) dt = 1 \quad (7)$$

when $t_0 > T$.

This is as much of the elementary pulse method as we shall need in order to reconsider the cases treated in Section 10.2. For the treatment of more general cases the method is described in greater detail in Appendix B.

The minimization of the integral (6) under the restriction (7) reduces to a simple isoperimetric problem in the calculus of variations, in cases in which $S(p)$ is a polynomial in p . It is essential first of all, however, to note that if $S(p)$ is of degree n , the integral (6) will converge only if $W(t)$ is differentiable at least n times. In other words, $W(t)$ must have continuous derivatives of all orders up to the $(n-1)$ th inclusive, although the n th derivative may have finite discontinuities. In particular, if $W(t)$ is to be zero outside of $0 \leq t \leq T$, its

derivatives of orders up to the $(n-1)$ th inclusive must vanish at both $t = 0$ and $t = T$. These $2n$ boundary conditions must be imposed on the solution of the Euler equation which in this case is

$$S\left(\frac{d}{dt}\right) \cdot S\left(-\frac{d}{dt}\right) \cdot W(t) = \lambda.$$

λ is a constant parameter which is finally adjusted to that the restriction (7) is satisfied.

The first case treated in Section 10.2 is one in which $N(\omega^2) = 1$, whence $S(p) = 1$ and $F(t) = W(t)$. The integral (6) is a minimum under the restriction (7) if $W(t)$ is constant by intervals. The restriction (7) then requires $W(t)$ to be of the form (2).

The case of first derivative smoothing treated in 10.2 is one in which $N(\omega^2) = \omega^2$, whence $S(p) = p$ and $F(t) = \dot{W}(t)$. If the integral (6) is to converge at all, $\dot{W}(t)$ must not have discontinuities of impulsive or higher type; in other words, $W(t)$ must be continuous through all values of t . The integral is a minimum under the restriction (7) if $\dot{W}(t)$ is constant by intervals. The restriction (7) then requires $W(t)$ to be of the form (4).

These results may be generalized immediately. In whatever way the signal to be smoothed may have been derived from the tracking data, let the power spectrum of the noise associated with it be $N(\omega^2) = \omega^{2n}$. Then $S(p) = p^n$ and $F(t) = W^{(n)}(t)$. If the integral

(6) is to converge at all, $w^{(n-1)}(t)$ must be continuous through all values of t . The integral is a minimum under the restriction (7) if $W^{(2n)}(t)$ is constant by intervals. The restriction (7) then requires $W(t)$ to be of the form

$$W(t) = \frac{(2n+1)!}{(n!)^2 T} \left[\frac{t}{T} \left(1 - \frac{t}{T} \right) \right]^n \quad 0 < t < T. \quad (8)$$

It may be noted that the convergence requirements which arise in the foregoing discussion are directly related to the discussion and theorem in Section A.8, Appendix A, with respect to the relationship between discontinuities in the impulsive admittance and its derivatives on the one hand, and the ultimate cutoff characteristic of the transmission function on the other hand. The continuity of $W^{(n-1)}(t)$ is obviously required to make the transmission fall off ultimately at the rate of $6(n+1)$ db per octave against the rise of $6n$ db per octave in the noise power spectrum.

The integral (6) may also be used to evaluate the relative advantage of the best weighting function over another weighting function. As an example, consider the case where the weighting function (2) is the best. The value of the integral (6) in this case is $1/T$. If the weighting function (4) is used against the same noise, the value of the integral (6) is $6/5T$. Hence, as far as rms error or standard deviation is concerned, the second weighting function is $\sqrt{5/6}$ or 0.913 as efficient as the first.

SMOOTHING FUNCTIONS FOR GENERAL POLYNOMIAL EXPANSIONS

THE THEORY of "smoothing a constant" developed in the preceding chapter will be extended in this chapter to the problem of smoothing a polynomial function of time of any prescribed degree. The extension is, however, restricted to the case of a flat noise spectrum. In addition to the smoothing problem, the analysis also provides a way of designing a network which will extrapolate the polynomial a given distance t_f into the future. The network is so arranged that t_f is continuously variable. In addition, the degree of the polynomial can readily be changed to fit changes in the complexity of the assumed form of the data, apart from noise.

It is clear that these results amount, in a certain sense, to an alternative to Wiener's method for the design of prediction circuits for general time series. Thus, to predict a time series of any given complexity we would need only to begin with a polynomial of sufficiently high degree to fit the observed data, and extrapolate. Aside from the restriction to a flat noise spectrum, perhaps the most obvious difference from Wiener's method is the fact that the settling time restriction limits the data upon which the prediction rests to a finite interval in the past. To advance such a prediction theory seriously, however, it would be necessary to go much farther into the way in which the degree of the polynomial is established and the justification for assuming that the extrapolated value represents a probable future value for the function.*

This general discussion will not be undertaken here. Since prediction with high degree polynomials will certainly be sensitive to minor irregularities in the data, tracking errors would necessarily limit the application of the method in any case. If we confine ourselves to reasonably low degree polynomials, however,

* As an example of possible difficulties we may notice the fact that two polynomials of different degree which approximate a given function as closely as possible, in a least squares sense, in a prescribed interval frequently differ radically outside that interval.

the method is useful. An example is furnished by the prediction of airplane position, in rectangular coordinates, by quadratic functions of time. Here the square terms represent the effects of accelerations in the various coordinates. We can defend the inclusion of such terms on the ground that it is plausible to assume that an airplane may experience constant accelerations, due to turns, the force of gravity, etc., for considerable periods of time. The linear term represents plane velocity and needs no defense. The constant term, of course, gives the plane position at some reference time. Including it in the smoothing operation is equivalent to introducing "present-position" smoothing of the sort suggested by the broken lines in Figure 1 of Chapter 7.^b

Aside from its direct interest as a possible prediction method, the analysis in this chapter is also of indirect interest for the additional light it sheds on the effect of the noise spectrum on smoothing functions. It turns out that smoothing a power of time, with a flat noise spectrum, is equivalent to smoothing a constant with a somewhat different noise spectrum. Thus the smoothing functions developed for polynomials are also useful as special cases of smoothing functions applicable to constants.

11.1

GENERAL METHOD

Let λ be any past value of time and let t be the present value. If the data is fitted with a smooth curve $\bar{E}(\lambda)$, the predicted value may be taken as $\bar{E}(t + t_f)$. The procedure of fitting is the familiar one of minimizing the integral

$$\int_{-\infty}^t [\bar{E}(\lambda) - E(\lambda)]^2 W_0(t, \lambda) d\lambda$$

^b In the circuit of Figure 1, Chapter 7, however, the smoothing network would produce a lag in the present-position data delivered to the prediction circuit, and this lag would, of course, mean some error in following a moving target. In the method described in this chapter such lags are automatically compensated for by adjustments in the coefficients of the other terms of the polynomial.

with respect to disposable parameters in $\bar{E}(\lambda)$ and a prescribed weighting function $W_0(t, \lambda)$. The lower limit of the integral is indicated as $-\infty$ in compliance with the physical impossibility of discriminating between relevant and irrelevant data, with fixed linear networks, except on the basis of age. The burden of discrimination must be relegated to the weighting function which must be a function only of the age $t - \lambda$. Under the ideal restriction that $W_0(t - \lambda)$ is identically zero when $t - \lambda > T$ or $\lambda < t - T$, the indicated lower limit of the integral is purely nominal.

As in Section 10.2, it is convenient to conduct the analysis in terms of the age variable $\tau = t - \lambda$ introduced there. If

$$\bar{F}(\tau) = \bar{E}(\lambda) \quad F(\tau) = E(\lambda)$$

the integral to be minimized may be expressed in the form

$$\int_0^\infty [\bar{F}(\tau) - F(\tau)]^2 W_0(\tau) d\tau. \quad (1)$$

In accordance with the discussion of quasi-distortionless transmission networks in Section A.10, Appendix A, the smooth curve $E(\lambda)$ should be a polynomial in λ . Hence $\bar{F}(\tau)$ should be a polynomial in τ . It will be more convenient, however, to express $F(\tau)$ formally as a linear combination of polynomials in τ which may be orthogonalized. Hence, let

$$\bar{F}(\tau) = V_0 + V_1 \cdot G_1(\tau) + V_2 \cdot G_2(\tau) + \dots + V_n \cdot G_n(\tau) \quad (2)$$

where $G_m(\tau)$ is an m th degree polynomial in τ .

Let $W_0(\tau)$ be normalized in the sense that

$$\int_0^\infty W_0(\tau) d\tau = 1$$

and the $G_m(\tau)$ be orthogonalized with respect to the weighting function $W_0(\tau)$ in the sense that

$$\int_0^\infty G_l(\tau) G_m(\tau) W_0(\tau) d\tau = \begin{cases} 0 & \text{if } l \neq m \\ \frac{1}{k_m} & \text{if } l = m \end{cases}$$

$$(G_0 = 1, k_0 = 1).$$

The integral (1) is then a minimum with respect to the V_m 's in (2) if

$$V_m = k_m \int_0^\infty F(\tau) \cdot G_m(\tau) \cdot W_0(\tau) d\tau. \quad (3)$$

In terms of the forward time λ , (2) and (3) reduce to

$$\bar{E}(\lambda) = V_0(t) + V_1(t) \cdot G_1(t - \lambda) + V_2(t) \cdot G_2(t - \lambda) + \dots + V_n(t) \cdot G_n(t - \lambda) \quad (4)$$

where

$$V_m(t) = k_m \int_{-\infty}^t E(\lambda) \cdot G_m(t - \lambda) \cdot W_0(t - \lambda) d\lambda. \quad (5)$$

Expression (5) identifies the $V_m(t)$ as the responses to $E(\lambda)$ of fixed linear networks whose impulsive admittances are

$$W_m(\tau) = k_m G_m(\tau) : W_0(\tau). \quad (6)$$

By (4), the predicted value may be obtained by a linear combination of the responses of these networks, viz.,

$$\bar{E}(t + t_f) = V_0(t) + G_1(-t_f) \cdot V_1(t) + G_2(-t_f) \cdot V_2(t) + \dots + G_n(-t_f) \cdot V_n(t). \quad (7)$$

A schematic representation of an n th order smoothing and prediction circuit, based on (7), is shown in Figure 1, where the $G_m(-t_f)$ are represented as potentiometer factors dependent on the time of flight.

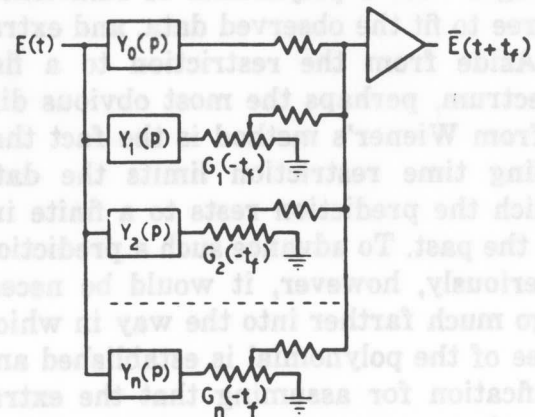


FIGURE 1. Schematic representation of n th order smoothing and prediction circuit.

Alternatively, (7) may be written

$$\bar{E}(t + t_f) = \bar{E}(t) + [G_1(-t_f) - G_1(0)] \cdot V_1(t) + \dots + [G_n(-t_f) - G_n(0)] \cdot V_n(t) \quad (8)$$

where $\bar{E}(t)$ is then replaced by $E(t)$ when position data smoothing is to be omitted.

It is not necessary that the $G_m(\tau)$ polynomials be orthogonal. However, the circuit switching required to reduce or increase the order of the prediction is simplest when the $G_m(\tau)$ polynomials are orthogonal. Orthogonal polynomials corresponding to any prescribed

weighting function $W_0(\tau)$ are readily derived by well-known methods.

The weighting function $W_0(\tau)$ may be determined by either of the methods described in Appendix B as the best weighting function for smoothing position data, under prescribed tracking error characteristics. Then the best impulsive admittances $W_m(\tau)$ for a smoothing and prediction circuit, are prescribed by (6).

The relationship (6) shows that if the prescribed weighting function $W_0(\tau)$ satisfies the formal requirements for physical realizability, so will all of the impulsive admittances $W_m(\tau)$. Of the standard sets of orthogonal polynomials those of Laguerre appear to be the best adapted to physical realization. The Laguerre polynomials $L_n^{(\alpha)}(\tau)$ are orthogonal in $0 \leq \tau < \infty$ with the weighting function $\tau^\alpha e^{-\tau}$. However, such a weighting function is, in general, very unsatisfactory from the practical point of view of settling characteristics.

It is possible of course to approximate any prescribed weighting function $W_0(\tau)$ as closely as may be desired in a physically realizable form, derive a set of orthogonal polynomials based on the approximate form, and determine the impulsive admittances $W_m(\tau)$ from (6). However, such a procedure leads to complexities of network configuration which increase very rapidly with the index m . This increasing complexity is hardly justifiable in practice.

From the foregoing considerations, it appears that the most practical procedure is to derive all of the impulsive admittances $W_m(\tau)$ without regard to physical realizability, approximate them independently in physically realizable forms of independently prescribed complexities, and modify or redetermine the potentiometer factors in accordance with the discussion in Section A.10, Appendix A.

11.2 WEIGHTING FUNCTIONS FOR DERIVATIVES

The impulsive admittances defined by (6) for $m > 0$ may not be regarded as weighting functions even though the response of the corresponding networks to $E(\lambda)$ is, by (5)

$$V_m(t) = \int_0^\infty E(t - \tau) \cdot W_m(\tau) \cdot d\tau, \quad (10)$$

because, with the exception of $W_0(\tau)$, the $W_m(\tau)$, as will presently be seen, cannot be normalized. The term weighting function is reserved for the functions defined by (11) below.

Since τ^r is a linear combination of the $G_s(\tau)$ where $s = 0, 1, \dots, r$, it is obvious from (6) that

$$\int_0^\infty \tau^r W_m(\tau) d\tau = 0$$

when $r < m$.

In particular

$$\int_0^\infty W_m(\tau) d\tau = 0$$

when $m > 0$.

Since the transmission function $Y_m(p)$ of a network is the Laplace transform of its impulsive admittance (see Section A.3), we have

$$\begin{aligned} Y_m(p) &= \int_0^\infty W_m(\tau) e^{-p\tau} d\tau \\ &= \sum_{r=0}^\infty \frac{(-p)^r}{r!} \int_0^\infty \tau^r W_m(\tau) d\tau. \end{aligned} \quad (9)$$

The first m terms in this series vanish. Hence $Y_m(p)$ will be of the form

$$Y_m(p) = p^m y_m(p) \quad (10)$$

where $y_m(0) \neq 0$. This permits us to regard the network whose impulsive admittance is $W_m(\tau)$ as an instantaneous m th order differentiator, corresponding to the factor p^m in (10), in tandem with a purely smoothing network whose transmission function is $y_m(p)$.

It is convenient to associate a weighting function $w_m(\tau)$ with the purely smoothing network whose transmission function is $y_m(p)$. Dividing (10) through by p^m the resulting operational equation may be interpreted (see Section A.5) to mean that the weighting function $w_m(\tau)$ is the m -fold integral of the impulsive admittance $W_m(\tau)$ between the limits 0 and τ . This is expressed by

$$w_m(\tau) = \int_0^\tau \cdots \int_0^\tau W_m(\tau) \cdot (d\tau)^m. \quad (11)$$

By a relationship similar to (9) between $y_m(p)$ and $w_m(\tau)$, it follows from $y_m(0) \neq 0$ that

$$\int_0^\infty w_m(\tau) d\tau \neq 0.$$

Hence the $w_m(\tau)$ may be normalized in the sense that

$$\int_0^\infty w_m(\tau) d\tau = 1$$

for all values of m . However, this may be done in general only if the $G_m(\tau)$ polynomials are not normalized in the sense that $k_m = 1$ for any value of $m > 0$. It is in fact readily shown that the coefficient of τ^m in $G_m(\tau)$ must be the same as that of τ^m in $e^{-\tau}$.

11.3 LEGENDRE POLYNOMIALS

The Legendre polynomials $P_m(x)$ are orthogonal with respect to the range $-1 \leq x \leq 1$ and uniform weighting. In other words, the polynomials $P_m(2\tau - 1)$ are orthogonal with respect to the range $0 \leq \tau \leq \infty$ and the weighting function^c

$$W_0(\tau) = 1 \quad \text{when } 0 \leq \tau \leq 1 \\ = 0 \quad \text{when } \tau > 1.$$

It is known from Section 10.4 that this form for the weighting function $W_0(\tau)$ is best in case the tracking errors are flat random noise. In the integral (1) to be minimized, the $G_m(\tau)$ polynomials should then be

$$G_m(\tau) = (-1)^m \frac{m!}{(2m)!} P_m(2\tau - 1).$$

The first few of these are tabulated below.

m	$G_m(\tau)$
0	1
1	$\frac{1}{2} - \tau$
2	$\frac{1}{12} - \frac{\tau}{2} + \frac{\tau^2}{2}$
3	$\frac{1}{120} - \frac{\tau}{10} + \frac{\tau^2}{4} - \frac{\tau^3}{6}$

With the help of the formula

$$\int_{-1}^1 [P_m(x)]^2 dx = \frac{2}{2m+1}$$

^c The unit of time being equal to the nominal smoothing time.

it is readily determined that

$$\frac{1}{k_m} = \int_0^\infty [G_m(\tau)]^2 W_0(\tau) d\tau \\ = \frac{(m!)^2}{(2m)!(2m+1)!}.$$

Then, by (6)

$$W_m(\tau) = (-1)^m \frac{(2m+1)!}{m!} P_m(2\tau - 1) \quad 0 \leq \tau \leq 1 \\ = 0 \quad \tau > 1.$$

Substituting this in turn into (11) and making use of Rodrigues' formula

$$P_m(x) = \frac{(-1)^m}{2^m m!} \frac{d^m}{dx^m} (1-x^2)^m$$

or

$$P_m(2\tau - 1) = \frac{(-1)^m}{m!} \frac{d^m}{d\tau^m} [\tau(1-\tau)]^m$$

it is finally found that

$$w_m(\tau) = \frac{(2m+1)!}{(m!)^2} [\tau(1-\tau)]^m \quad 0 \leq \tau \leq 1 \\ = 0 \quad \tau > 1. \quad (12)$$

By a relationship of the form of (9) the transmission functions $y_m(p)$ corresponding to the weighting functions $w_m(\tau)$ may be determined. The first three are

$$y_0(p) = \frac{1 - e^{-p}}{p} \\ y_1(p) = \frac{6}{p^3} [(p-2) + (p+2)e^{-p}] \\ y_2(p) = \frac{60}{p^5} [(p^3 - 6p + 12) - (p^3 + 6p + 12)e^{-p}].$$

These may be written in the form

$$y_m(p) = Q_m(\omega) \cdot e^{-i\omega/2} \quad (13)$$

where

$$Q_0(\omega) = \frac{\sin x}{x} \quad \left(x = \frac{\omega}{2}\right) \\ Q_1(\omega) = 3 \frac{\sin x - x \cos x}{x^3} \\ Q_2(\omega) = 15 \frac{(3 - x^2) \sin x - 3x \cos x}{x^5} \quad (14)$$

or in the infinite power-series form

$$y_0(p) = \sum_{n=0}^{\infty} \frac{(-p)^n}{(n+1)!}$$

$$y_1(p) = 6 \sum_{n=0}^{\infty} \frac{n+1}{(n+3)!} (-p)^n$$

$$y_2(p) = 60 \sum_{n=0}^{\infty} \frac{(n+1)(n+2)}{(n+5)!} (-p)^n. \quad (15)$$

Methods for obtaining physically realizable approximations to the weighting functions $w_m(\tau)$ or impulsive admittances $W_m(\tau)$, based upon the Q functions (14) and the series expansions (15) are described in Chapter 12.

PHYSICAL REALIZATION OF DATA-SMOOTHING FUNCTIONS

THIS CHAPTER will be devoted to a brief review of some of the methods and techniques which have been used in the physical realization of data-smoothing or weighting functions. The first two sections will be devoted to methods for determining physically realizable approximations to a desired weighting function. The third section takes up the use of feedback amplifiers and servomechanisms in order to avoid the use of coils of generally fantastic sizes. The final section takes up the design of resistance-capacitance networks.

Methods of deriving physically realizable approximations of best weighting functions may be divided into two classes, which may be called, for convenience, *t*-methods and *p*-methods. The *t*-methods are those in which a prescribed best weighting function $W(t)$ is approximated directly by a function $W_a(t)$ of realizable form, viz., a sum of decaying exponential terms and exponentially decaying sinusoidal terms. However, the *t*-methods are most useful when the approximation is restricted to a sum only of exponential terms. According to the discussion in Section A.9, Appendix A, such a restriction corresponds physically to passive *RC* transmission networks. A *t*-method was used by Phillips and Weiss in the reference quoted in Section 10.3 to obtain an approximation with one decaying exponential term and one exponentially decaying sinusoidal term. However, this method rapidly becomes unwieldy as the number of terms is increased.

The *p*-methods are those in which the approximation is derived indirectly from the transmission function $Y(p)$ corresponding to $W(t)$. A rational function $Y_a(p)$ approximating $Y(p)$ is first determined. If it is realizable, and it usually is, then $W_a(t) = L^{-1}[Y_a(p)]$. In general, $Y_a(p)$ will have complex poles and, therefore, $W_a(t)$ will have exponentially decaying sinusoids as well as simple exponentials. This gives the *p*-methods a considerable advantage over the *t*-methods in more efficient use of network elements. The fact that this generally calls for impractical element values in passive

RLC networks is not serious. As shown in Section 12.3, the use of coils may be avoided entirely by the use of feedback amplifiers.

12.1

t-METHODS

To describe the *t*-method,^a let

$$W_a(t) = A_1 e^{-\alpha_1 t} + A_2 e^{-\alpha_2 t} + \dots + A_n e^{-\alpha_n t} \quad (1)$$

where the α 's are prescribed and the A 's are to be determined. Two considerations are involved in the determination of the A 's. The first consideration is based on the relationship between the continuity conditions at $t = 0$ and the ultimate slope of the loss characteristic as expressed in the theorem in Section A.8. Accordingly, a number of relations of the type

$$\begin{aligned} A_1 + A_2 + \dots + A_n &= 0 \\ \alpha_1 A_1 + \alpha_2 A_2 + \dots + \alpha_n A_n &= 0 \end{aligned} \quad (2)$$

$$\alpha_1^r A_1 + \alpha_2^r A_2 + \dots + \alpha_n^r A_n = 0 \quad r < n - 1$$

must be satisfied. This leaves $n - r - 1$ of the A 's for the second consideration.

The second consideration concerns the manner in which the approximation in the range $t > 0$ is to be made. The approximation may, for example, be required to pass through $n - r - 1$ points on $W(t)$ or, the first $n - r - 1$ moments of the approximation may be required to be equal to the corresponding moments of $W(t)$. The latter is expressed by relations of the type

$$\begin{aligned} \frac{A_1}{\alpha_1^s} + \frac{A_2}{\alpha_2^s} + \dots + \frac{A_n}{\alpha_n^s} &= \frac{1}{(s-1)!} \int_0^\infty W(t) t^{s-1} dt \\ s &= 1, 2, \dots, n - r - 1 \end{aligned} \quad (3)$$

Foster's investigations were concerned only with the parabolic weighting function (4) Chapter 10, so that only the first of (2) was involved. Numerical studies led to the belief that, with a given number of α 's, the best approximation was to be had from the case in

^a The *t*-method is principally due to R. M. Foster.

which all of the α 's are equal. Hence the natural center of attention was the special form

$$W_a(t) = (A_1 t + A_2 t^2 + \dots + A_{n-1} t^{n-1}) e^{-\alpha t}. \quad (4)$$

At large values of t this expression reduces approximately to the last term, and if it is assumed that $A_{n-1} \doteq 1$, the settling condition fixes α to at least a first approximation. The rest of the work of approximating the parabola is then equivalent to a problem in polynomial approximation. Once the A 's are determined, a better value of α can be found from the settling condition, and the process gone through again.

If the α 's are only approximately equal, the approximation will still behave approximately like (4) with an average value used for α . The difficulty with equal or nearly equal α 's is that it leads to networks with extreme element values. In order to secure satisfactory element values, it is generally necessary to depart substantially from the condition of equal α 's. This results in some, but not a large, loss of efficiency in approximating the parabola. Foster recommends that the α 's be chosen as a geometric series, with their geometric mean more or less around the equivalent point for equal α 's. With four α 's he suggests that the constant ratio in the series may be 3:2, whereas with only two α 's the ratio should be raised to 2:1. These are, however, only rough values and obviously depend on individual opinion of what constitutes an unreasonable element value.

As a matter of experience, it turns out that the characteristic first obtained usually has a rather long and slowly decaying tail, as shown in Figure 1. This, of course, is equivalent to a

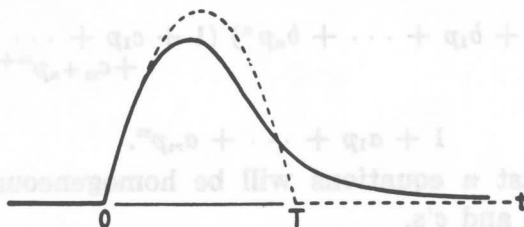


FIGURE 1. Approximation to parabolic weighting function, showing poor settling characteristic.

correspondingly long "settling time," or time before a useful prediction can be made. In practice, therefore, after the preliminary design has been found, adjustments are made to bring the tail of the curve under control,

partly by modifying the values of the A 's slightly, and partly by contracting the time scale to bring the part of the tail which remains appreciable within the allowable settling time limits. This leads to the somewhat lopsided match to the parabola shown in Figure 2.

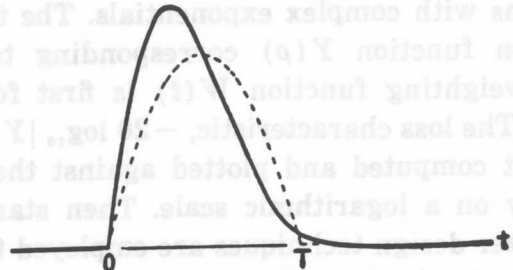


FIGURE 2. Approximation to parabolic weighting function, showing better settling characteristic.

A method of bringing the tail of the curve under control^b is to minimize the expression

$$\int_0^\infty [W_a(t)]^2 dt = \sum_{l,m=1}^n C_{lm} A_l A_m \quad (5)$$

where

$$C_{lm} = \frac{e^{-(\alpha_l + \alpha_m)T}}{\alpha_l + \alpha_m}$$

under the restrictions (2) and all but the last of (3).

The t -method used by Phillips and Weiss is based on a 3-term approximation of the form (1) in which one α is real while the other two may be conjugate complex. The α 's are not prescribed, so that there are six parameters to be determined. Four restrictions are imposed, viz., the first of (2), the first of (3), a restriction on the value of the tail area, viz.,

$$\int_T^\infty W_a(t) dt = \sum_{l=1}^3 \frac{A_l e^{-\alpha_l T}}{\alpha_l},$$

and the cross-over condition

$$W_a(T) = 0.$$

Finally, the transmitted noise power, which, under the assumption of flat random noise associated with the position data, takes the form (see Section 10.4)

$$\int_0^\infty [W_a(t)]^2 dt$$

is minimized with respect to the two remaining parameters by numerical methods.

^b Used by R. F. Wick.

12.2

p-METHODS

Three p -methods have been used. These will be described in chronological order.

The first p -method is one which was used by R. L. Dietzold in exploiting the use of feedback amplifiers to secure the advantages of approximations with complex exponentials. The transmission function $Y(p)$ corresponding to the best weighting function $W(t)$ is first formulated. The loss characteristic, $-20 \log_{10} |Y(i\omega)|$, is next computed and plotted against the frequency on a logarithmic scale. Then standard equalizer design techniques are employed to approximate the loss characteristic, keeping in mind that the transmission loss in the feedback network of a feedback amplifier becomes a transmission gain for the circuit as a whole (see Section 12.3).

The second p -method is merely a more complete analytic formulation of the first, thereby avoiding the necessity for employing equalizer design techniques. It depends upon the possibility of expressing the transmission function corresponding to the best weighting function, in the form of equation (13) Chapter 11, which is associated with the symmetry of the weighting function, as shown in Section A.7. The method is based upon the determination of the envelope of the Q -function. The Q -function is first differentiated in order to obtain the equation which determines the values of ω at which the maxima and minima occur. This transcendental equation is not solved but is used to eliminate the trigonometric functions in the expression of the Q -function. The resulting expression, which is an irrational function of ω^2 , is then squared in order to make it a rational function of ω^2 . The substitution $p^2 = -\omega^2$ is made and the expression is then resolved into two factors of which one contains all the poles with negative real parts while the other contains all the poles with positive real parts, the two factors being conjugate complex when $p = i\omega$. The first factor is then taken as an approximation of the desired transmission function. Applying the method to the desired transmission functions defined by (13) and

(14) of Chapter 11, we get

$$\begin{aligned} y_0(p) &\doteq \frac{2}{2+p} \\ y_1(p) &\doteq \frac{12}{12+6p+p^2} \\ y_2(p) &\doteq \frac{120}{120+60p+12p^2+p^3}. \end{aligned} \quad (6)$$

This last is the basis for the design of a position and rate smoothing circuit for a proposed computer for controlling bombers from the ground.^{11,12} This design is described briefly in Chapter 13.

The third p -method is based upon the ascending power-series expansion of the transmission function corresponding to the best weighting function. Examples of such power series are given by (15) of Chapter 11. The method of approximation is one which is credited to Padé in O. Perron's "Kettenbrüchen."¹³ If the discussion in Section A.8 is referred to, it will be seen to be also a method of moments.

The method consists in determining the coefficients in a rational function of the form

$$\frac{1 + a_1p + a_2p^2 + \dots + a_mp^m}{1 + b_1p + b_2p^2 + \dots + b_np^n} \quad (7)$$

so that the ascending power-series expansion of the rational function will agree with that of the best transmission function, term for term up to and including p^{m+n} . If the series for the best transmission function is

$$1 + c_1p + c_2p^2 + \dots + c_{m+n}p^{m+n} + \dots \quad (8)$$

the equations which determine the coefficients in (7) are obtained by equating coefficients of corresponding powers of p , up to and including the $(m+n)$ th, in

$$(1 + b_1p + \dots + b_np^n)(1 + c_1p + \dots + c_{m+n}p^{m+n})$$

and

$$1 + a_1p + \dots + a_mp^m.$$

The last n equations will be homogeneous in the b 's and c 's.

It has been expedient in some cases to omit the last few of the $(m+n)$ equations in order to have some control over the number of real roots and poles and the number of conjugate pairs of complex roots and poles in the resulting rational function.

In the assumed rational expression (7) the

difference $n - m$ should be chosen so that the ultimate slope of the loss characteristic will be the same as for the best transmission function. According to the theorem in Section A.8, if $W(t)$ behaves like t^r as $t \rightarrow 0$, we should take $n - m = r + 1$. As a matter of experience the rational expression has invariably turned out to be physically realizable whenever this "rule" was followed. Frequently, however, the rational expression has turned out to be physically realizable under small departures from the rule.

Examples of this method are given in Chapter 13.

13.3 USE OF FEEDBACK AMPLIFIERS AND SERVOMECHANISMS

In this section we shall describe the use of feedback amplifiers and servomechanisms to obtain desired transmission functions. For complete discussions of the most recent technical advances in the analysis and design of feedback amplifiers and servomechanisms the reader should consult some of the modern literature on these subjects.^{2,3,5,15,16,17}

Let us assume that we have two networks whose transmission functions are $Y_1(p)$ and $Y_2(p)$, respectively, as shown in Figure 3. For

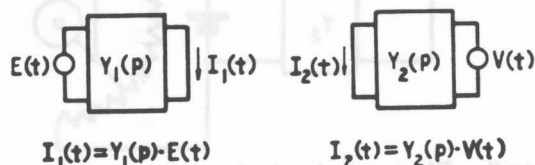


FIGURE 3. Schematic representation of networks intended for feedback circuit application.

a signal $E(t)$ applied to the first network the short-circuit output current is $I_1(t) = Y_1(p) \cdot E(t)$. For a signal $V(t)$ applied to the second network the short-circuit output current is

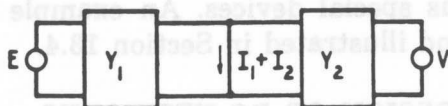


FIGURE 4. First step in combining networks.

$I_2(t) = Y_2(p) \cdot V(t)$. With the networks sharing a common short-circuiting conductor as shown in Figure 4, the current through the conductor is $I_1 + I_2$. If the source which develops the volt-

age $V(t)$ across the input terminals of the second network were in fact under the control of the current through the conductor, as shown schematically in Figure 5, in such a manner

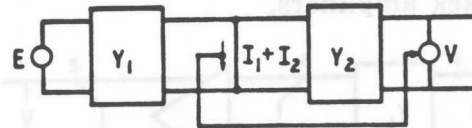


FIGURE 5. Output voltage controlled by short-circuit current across intermediate terminals.

that it had to develop that voltage $V(t)$ which reduces the current in the conductor to zero, then

$$Y_1(p) E(t) + Y_2(p) \cdot V(t) = 0.$$

Hence, the transmission function (now a voltage-voltage ratio) of the arrangement shown in Figure 5 must be

$$Y(p) = -\frac{Y_1(p)}{Y_2(p)}. \quad (9)$$

This relationship provides a method of obtaining transmission functions with complex poles without the requirement of coils.^c The complex roots of $Y(p)$, must be assigned to the numerator of $Y_1(p)$, and the complex poles of $Y(p)$ to the denominator of $Y_2(p)$. Aside from this, the other roots and poles of $Y(p)$ may be assigned in any way which is favorable to good design practice. Redundant factors may be introduced if they are desirable, as is done in the examples described in Sections 13.1.5 and 13.3.

The source of the voltage $V(t)$ in Figure 5 does not have to be controlled by the current through the short-circuiting conductor. Since the current through any short circuit must be zero if the voltage across the short-circuited terminals is zero before the short circuit is connected across them, the source of the voltage $V(t)$ may just as well be controlled by the open-circuit voltage, as shown in Figure 6. It is clear that the source of the voltage $V(t)$ is ideally an infinite gain amplifier. It is not necessary, however, that the amplifier have ideally unilateral transmission and infinite input and output impedances, since departures from these ideal characteristics may be compensated for in the design of the feedback network.

The simple result expressed by (9) may be readily modified to take account of the finite

^c This observation was first made by R. L. Dietzold.

gain of a physical amplifier. The modification will be expressed as an extra factor which corresponds to the " $\mu\beta$ effect" or " $\mu\beta$ error"^{15e} commonly encountered in the theory and design of feedback amplifiers.

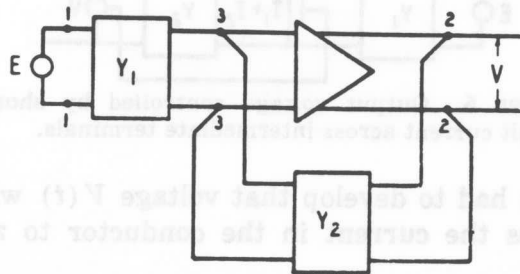


FIGURE 6. Output voltage controlled by open-circuit voltage across intermediate terminals.

The exact transmission function of the circuit shown in Figure 6 is most simply expressed in terms of the following quantities:

$Y_1(p)$ = current through a short across terminal-pair No. 3, per unit emf applied across terminal-pair No. 1.

$Y_2(p)$ = current through a short across terminal-pair No. 3, per unit emf applied across terminal-pair No. 2.

$Z_2(p)$ = impedance between terminal-pair No. 2, with terminal-pair No. 3 shorted.

$Z_3(p)$ = impedance between terminal-pair No. 3, with amplifier dead, terminal-pair No. 1 shorted, and terminal-pair No. 2 open.

$$G(p) = \text{transadmittance of amplifier.}$$

Then

$$Y = -\frac{Y_1}{Y_2} \frac{1 + \frac{Y_2}{G}}{1 - \frac{1}{GY_2Z_2Z_3}}. \quad (10)$$

The quantity $GY_2Z_2Z_3$ is the $\mu\beta$ of the circuit. The quantity $Y_1Y_2Z_2Z_3$ to which Y reduces when $G = 0$ represents the direct transmission of the circuit.

The active impedance across terminal-pair No. 2 is

$$Z_{2A} = \frac{Z_{2P}}{1 - GY_2 Z_2 Z_3} \quad (11)$$

where

$$Z_{2P} = Z_2(1 + Y_2^2 Z_2 Z_3) . \quad (12)$$

Z_{2p} is the passive impedance across terminal-pair No. 2. It differs from Z_2 in that terminal-pair No. 3 is open.

The exact expression (10) of the transmission function is useful chiefly as a check on the simpler but approximate expression (9). It is in general quite practicable to make the transadmittance or transconductance G of the amplifier large enough so that the $\mu\beta$ effect may be neglected.

In accordance with the sense in which the term "servomechanism" is used by MacColl,⁴ a feedback circuit, such as that shown in Figure 6, is a servomechanism — more specifically, an electronic servomechanism — since it operates on the ideal principle of maintaining zero voltage across the terminal-pair No. 3. An electromechanical counterpart of the circuit shown in Figure 6 is shown in Figure 7. These

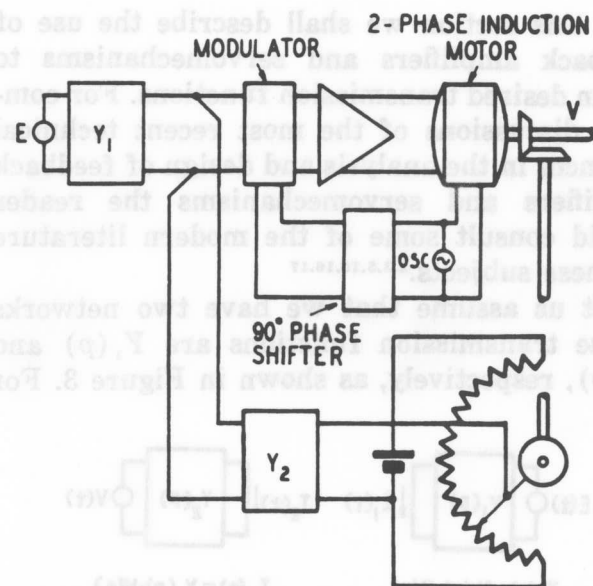


FIGURE 7. Electromechanical counterpart of feedback amplifier circuit resulting in servomechanism.

circuits assume that the signal $E(t)$ is a modulated d-c carrier.

If the signal is a modulated a-c carrier, "shaping" cannot be done conveniently by electrical networks. The difficulty may be avoided by various special devices. An example is described and illustrated in Section 13.4.

12.4 DESIGN OF RC NETWORKS

In this section we will describe and illustrate two general methods of designing *RC* networks. The first is most useful when the transmission function is finite and not zero at zero frequency; the second, when the transmission

function is zero at zero frequency. The case of a transmission function with a pole at zero frequency will not be considered, since it is covered by the methods described in the preceding section, in conjunction with the methods described below.

Let

$$Y(p) = \frac{a_0 + a_1p + \dots + a_{n+1}p^{n+1}}{1 + b_1p + \dots + b_np^n} \quad (a_0 > 0) \quad (13)$$

with simple, real, negative poles. Dividing by p , expanding into partial fractions and multiplying through by p , we get

$$Y(p) = \frac{pa_{n+1}}{b_n} + (a_0 + \frac{pA_1}{p + \alpha_1} + \frac{pA_2}{p + \alpha_2} + \dots) - \left(\frac{pB_1}{p + \beta_1} + \frac{pB_2}{p + \beta_2} + \dots \right)$$

where the A 's, B 's, α 's and β 's are positive real quantities. The first term must be associated with those in the first parentheses if $a_{n+1} > 0$, with those in the second parentheses if $a_{n+1} < 0$. The transmission function is now in the form

$$Y(p) = Y_A(p) - Y_B(p) \quad (14)$$

where $Y_A(p)$ and $Y_B(p)$ are physically realizable driving-point admittances of RC type. Each term of the form $pA/(p + \alpha)$ is the admittance of the two-terminal, two-element network

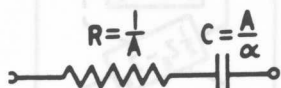


FIGURE 8. Simple RC network.

shown in Figure 8. Each term in (14) therefore represents a parallel combination of two-element networks of the type shown in Figure 8 and a conductance a_0 in the case of $Y_A(p)$,

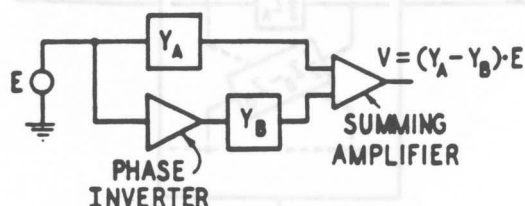


FIGURE 9. Method of realizing RC transmission functions, requiring phase inverter.

and a capacitance $|a_{n+1}|/b_n$ in the case of either $Y_A(p)$ or $Y_B(p)$. By well-known methods these two-terminal networks may be transformed into a variety of other configurations.

The transmission function (14) may be realized in the arrangement shown in Figure 9 or in that shown in Figure 10. The latter is a lattice network which is suitable only in a

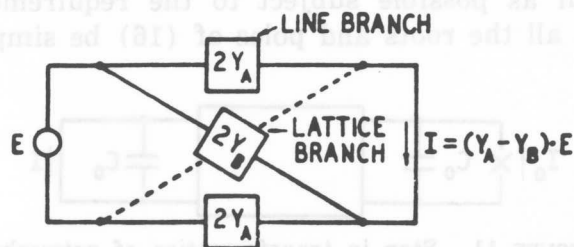


FIGURE 10. Lattice prototype for passive networks with RC transmission characteristics.

balanced-to-ground circuit. To obtain an unbalanced passive equivalent of this network we may resort to steps which will be described later in this section.

The second general method of designing RC networks is most useful when

$$Y(p) = p \frac{a_0 + a_1p + \dots + a_np^n}{1 + b_1p + \dots + b_np^n} \quad (a_0 > 0) \quad (15)$$

with simple, real, negative poles. Now, if the lattice in Figure 10 were driven from an infinite-impedance source of current I_0 , the output current would be

$$I = \frac{1 - \frac{Y_B}{Y_A}}{1 + \frac{Y_B}{Y_A}} I_0.$$

If, furthermore,

$$\frac{Y_B}{Y_A} = \frac{k - \frac{Y}{p}}{k + \frac{Y}{p}} \quad (16)$$

then

$$I = \frac{Y}{kp} I_0.$$

Taking it for granted for the moment that the lattice can be transformed as shown schematically in Figure 11, we may then discard the condenser across the output terminals and, by Thévenin's theorem,¹⁵ we may replace the condenser across the input terminals and the infinite-impedance current source by a series condenser and a zero-impedance voltage source. The result is shown in Figure 12. Since

$I_0 = pCE$ we now have

$$I = \frac{C_0}{k} Y \cdot E$$

which is the desired result, to a constant factor.

The factor k should in general be taken as small as possible subject to the requirement that all the roots and poles of (16) be simple,

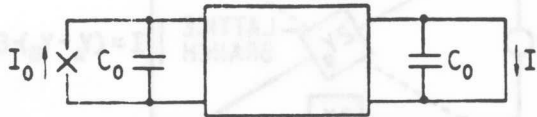


FIGURE 11. Step in transformation of networks with zero transmission at zero frequency.

real, and negative. It can always be taken large enough to fulfill this requirement. A suitable value may be easily chosen by inspection of a plot of $Y(p)/p$ for negative real values of p .

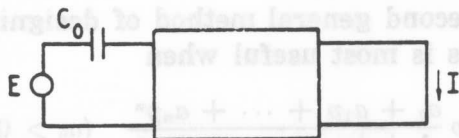


FIGURE 12. Final step in transformation of networks with zero transmission at zero frequency.

The numerator and denominator of (16) are of equal degree and therefore contain the same number of linear factors. These factors may be assigned to Y_A or to Y_B arbitrarily except that Y_A and Y_B must be physically realizable driving-point admittance functions which behave ultimately like condensers as the frequency increases indefinitely; that is, roots and poles must alternate and there must be a simple pole at infinity.

There are five kinds of steps which may be taken to transform a lattice into an unbalanced form. These steps are based upon Bartlett's bisection theorem,¹⁴ and may be taken in any order and as often as necessary. Each of them will now be described as it would be applied directly to Figure 10. In the following diagrams a lattice enclosed in a rectangle means an unbalanced network whose configuration may not be known yet, but whose lattice prototype is as indicated.

1. Shunt network pulled out of both branches: shown in Figure 13.

2. Shunt network pulled out of the line branch only: shown in Figure 14.

3. Series network pulled out of both branches: shown in Figure 15.^c

4. Series network pulled out of the lattice branch only: shown in Figure 16.^c

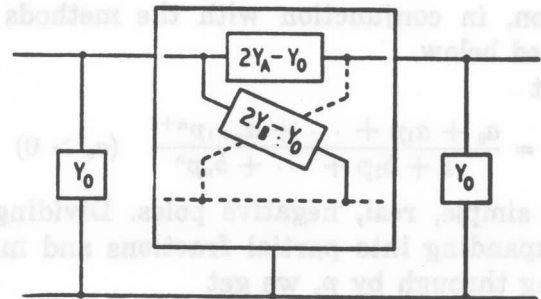


FIGURE 13. Step in transformation of lattice; shunt networks pulled out of both branches.

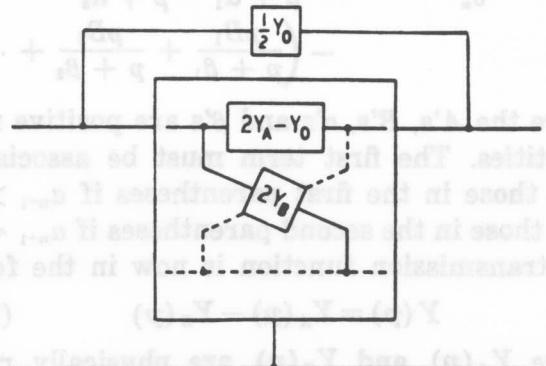


FIGURE 14. Step in transformation of lattice; shunt network pulled out of line branch only.

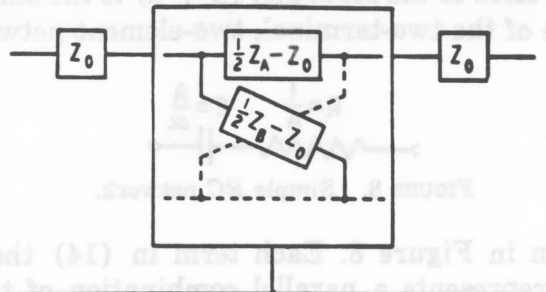


FIGURE 15. Step in transformation of lattice; series networks pulled out of both branches.

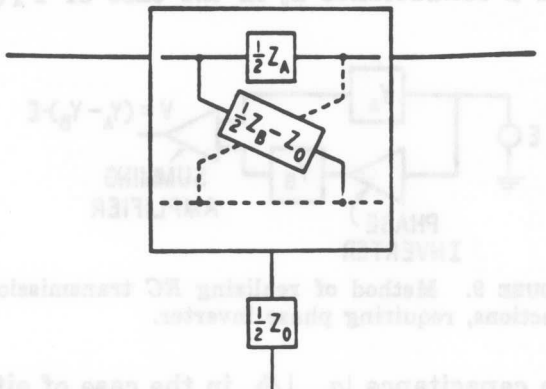


FIGURE 16. Step in transformation of lattice; series network pulled out of lattice branch only.

^c Given in impedance form.

5. Breakdown into parallel lattices: a fairly obvious step which need not be illustrated.

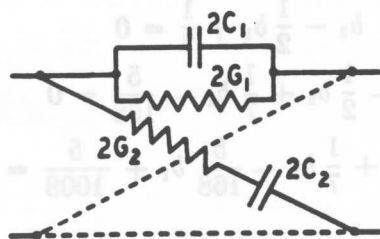
As an example of (13) consider

$$Y(p) = \frac{a_0 + a_1p + a_2p^2}{1 + b_1p}$$

where all the coefficients are positive. Since

$$Y(p) = \frac{pa_2}{b_1} + a_0 - \frac{(a_0b_1^2 - a_1b_1 + a_2)p}{b_1^2\left(p + \frac{1}{b_1}\right)}$$

there is no problem if $a_1 > (a_2/b_1) + a_0b_1$. But if $a_1 < (a_2/b_1) + a_0b_1$ we have the problem of trans-



$$\begin{aligned} C_1 &= \frac{a_2}{b_1} & G_1 &= a_0 \\ C_2 &= \frac{a_0b_1^2 - a_1b_1 + a_2}{b_1} \\ G_2 &= \frac{a_0b_1^2 - a_1b_1 + a_2}{b_1^2} \end{aligned}$$

FIGURE 17. Illustrative lattice prototype.

forming the lattice in Figure 17. We can apply steps 2 and 4 immediately, but find that the residual lattice cannot be transformed unless $a_1 > (a_2/b_1)$. Under this additional restriction we can apply step 3 obtaining finally the network shown in Figure 18.

As an example of (15) consider

$$Y(p) = p \frac{1 + 12p}{1 + 24p + 64p^2}.$$

Taking $k = 1$ (the smallest value which may be assigned), we get

$$\frac{Y_B}{Y_A} = \frac{2p(3 + 16p)}{(1 + 2p)(1 + 16p)}.$$

One way of choosing Y_A and Y_B is

$$Y_A = \frac{(1 + 2p)(1 + 16p)}{2(3 + 16p)} \quad Y_B = p.$$

This leads finally to the network shown in Figure 19. Such a simple network is possible of

course because $Y(p)$ happens to satisfy the requirements of a physically realizable driving-point admittance function. However, another way of choosing Y_A and Y_B is

$$Y_A = \frac{1 + 2p}{2} \quad Y_B = \frac{p(3 + 16p)}{1 + 16p}$$

This leads to the network shown in Figure 20.

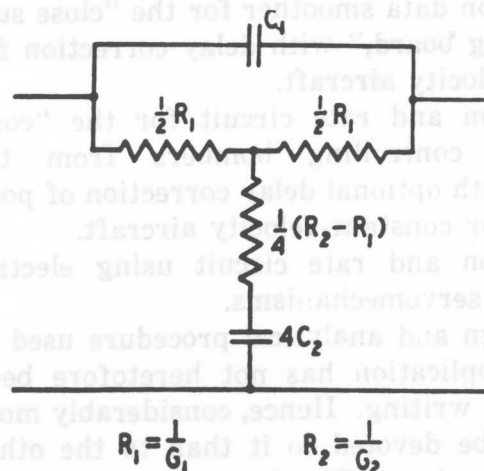


FIGURE 18. Unbalanced equivalent of illustrative lattice prototype when $a_2/b_1 < a_1 < (a_2/b_1) + a_0b_1$.

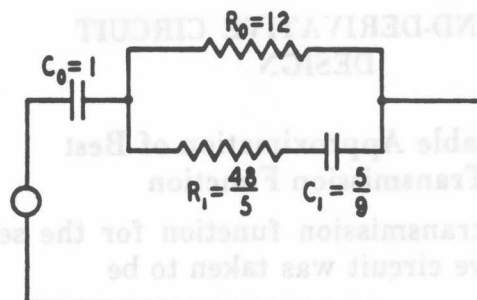


FIGURE 19. RC network with zero transmission at zero frequency.

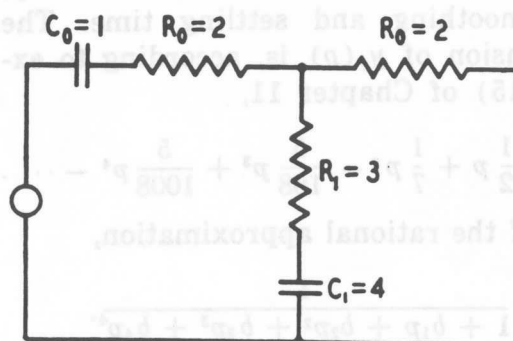


FIGURE 20. Another RC network with zero transmission at zero frequency.

ILLUSTRATIVE DESIGNS AND PERFORMANCE ANALYSIS

THE ILLUSTRATIVE MATERIAL described in this chapter is taken from four practical applications.

1. Second-derivative circuit for the M9 anti-aircraft director.

2. Position data smoother for the "close support plotting board," with delay correction for constant velocity aircraft.

3. Position and rate circuit for the "computer for controlling bombers from the ground," with optional delay correction of position data for constant-velocity aircraft.

4. Position and rate circuit using electro-mechanical servomechanisms.

The design and analytical procedure used in the first application has not heretofore been described in writing. Hence, considerably more space will be devoted to it than to the other three applications. The latter have been described in detail in reports.^{10,12,13}

13.1 SECOND-DERIVATIVE CIRCUIT DESIGN

13.1.1 Realizable Approximation of Best Transmission Function

The best transmission function for the second-derivative circuit was taken to be

$$Y_2(p) = p^2 y_2(p),$$

in the notation of Chapter 11. This assumes flat random noise in position data and, arbitrarily, 1-second smoothing and settling time. The series expansion of $y_2(p)$ is, according to expressions (15) of Chapter 11,

$$y_2(p) = 1 - \frac{1}{2}p + \frac{1}{7}p^2 - \frac{5}{168}p^3 + \frac{5}{1008}p^4 - \dots$$

The form of the rational approximation,

$$\bar{y}(p) = \frac{1}{1 + b_1 p + b_2 p^2 + b_3 p^3 + b_4 p^4},$$

was chosen for simplicity under the requirement that the transmission function $p^2 \bar{y}(p)$

should cut off at the rate of 12 db per octave.* This requirement was set as a precaution against noise due to granularity of the coordinate-conversion potentiometers in the director.

Following the procedure outlined in Section 12.2 the following equations were obtained:

$$b_1 - \frac{1}{2} = 0$$

$$b_2 - \frac{1}{2}b_1 + \frac{1}{7} = 0$$

$$b_3 - \frac{1}{2}b_2 + \frac{1}{7}b_1 - \frac{5}{168} = 0$$

$$b_4 - \frac{1}{2}b_3 + \frac{1}{7}b_2 - \frac{5}{168}b_1 + \frac{5}{1008} = 0$$

whence

$$b_1 = \frac{1}{2}, \quad b_2 = \frac{3}{28}, \quad b_3 = \frac{1}{84}, \quad b_4 = \frac{1}{1764}$$

Since

$$p^4 + 21p^3 + 189p^2 + 882p + 1764$$

$$= (p^2 + \frac{21 + \sqrt{21}}{2}p + 42)$$

$$\times (p^2 + \frac{21 - \sqrt{21}}{2}p + 42),$$

$\bar{y}_2(p)$ would have two conjugate pairs of complex poles, viz.,

$$p = -6.40 \pm i1.047, \quad -4.10 \pm i5.02,$$

of which one pair is very nearly real.

In order to simplify the circuit design, however, it was desirable to limit the number of complex poles to a single conjugate pair. This was accomplished by leaving b_4 arbitrary so that the denominator of $\bar{y}_2(p)$ was

$$1 + \frac{1}{2}p + \frac{3}{28}p^2 + \frac{1}{84}p^3 + b_4 p^4.$$

A value for b_4 which would make this expression vanish at two negative real values of p was found by plotting

$$1764b_4 = \frac{21}{x^4} (x^3 - 9x^2 + 42x - 84)$$

* The design antedated the formulation of the $n - m = r + 1$ rule given in Section 12.2, according to which the best transmission function should have been taken as $p^2 \bar{y}_2(p)$ in the notation of Chapter 11. However, no trouble was experienced in obtaining a physically realizable approximation, of the complexity assumed.

against x , as shown in Figure 1. The right-hand member is positive only in the range $x > 3.77$ and has a maximum of 0.982 at about $x = 6.63$.

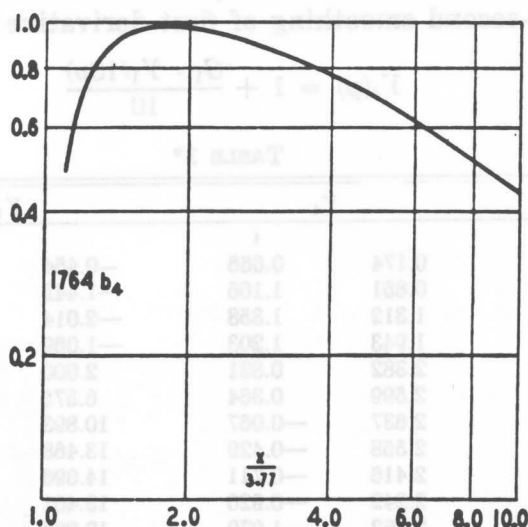


FIGURE 1. Graphical determination of b_4 .

In order to obtain a substantial separation between the two real poles of $\bar{y}_2(p)$, the value $1764b_4 = 0.5$ was chosen. The approximation

$$\bar{y}(p) = \frac{1}{1 + \frac{1}{2}p + \frac{3}{28}p^2 + \frac{1}{84}p^3 + \frac{1}{3528}p^4}$$

has poles at

$$p = -4.17391, -31.72813, -3.04898 \pm i 4.16463.$$

The series expansion of $\bar{y}_2(p)$ agrees with that of $y_2(p)$ to four terms, the fifth term being $37/7056 p^4$ instead of $5/1008 p^4$. The difference in the fifth term is less than 6 per cent.

The realized approximation and the best weighting function are shown in Figure 3.

13.1.2

Transient Responses

The responses of the physical network whose transmission function is $p^2\bar{y}_2(p)$ are compared to those of the best network whose transmission function is $p^2y_2(p)$, in Figures 2, 3, and 4. The signals for which (and the formulas by which) these responses were computed are tabulated below.

Figure	Signal		Response formulas	
	$t \leq 0$	$t \geq 0$	Realized	Best
2	0	1	$L^{-1}[p\bar{y}_2(p)]$	$60t(1-2t)(1-t)$
3	0	t	$L^{-1}[\bar{y}_2(p)]$	$30[t(1-t)]^2$
4	0	$\frac{1}{2}t^2$	$L^{-1}\left[\frac{1}{p}\bar{y}_2(p)\right]$	$t^3(10-15t+6t^2)$

It has been noted that Figure 3 also represents the best and the realized weighting functions.

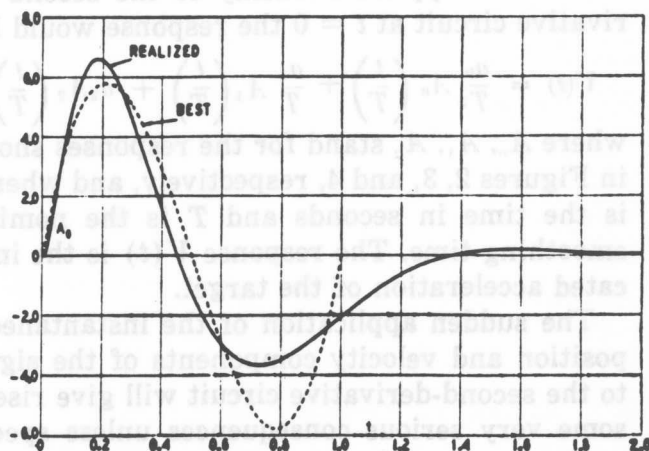


FIGURE 2. Responses to step function, viz., $E(t) = 1$ when $t > 0$.

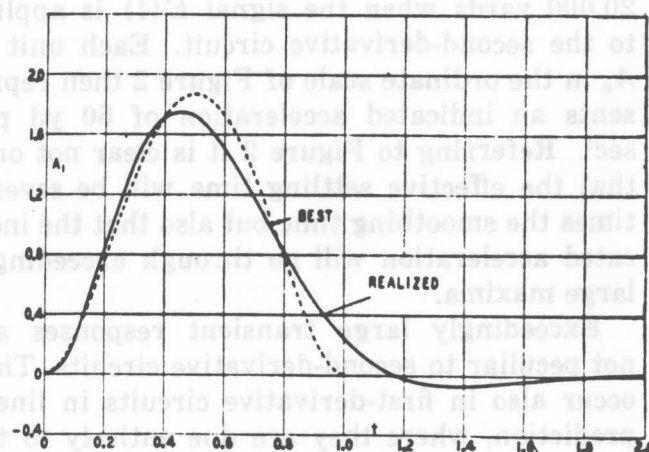


FIGURE 3. Responses to linear ramp function, viz., $E(t) = t$ when $t > 0$; second derivative smoothing functions.

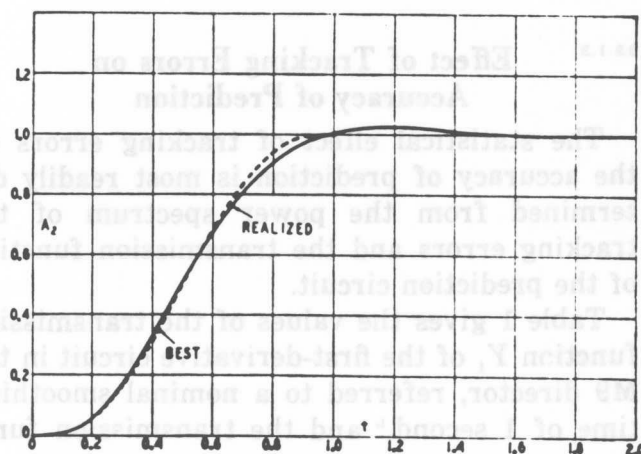


FIGURE 4. Responses to parabolic ramp function, viz., $E(t) = (\frac{1}{2})t^2$ when $t > 0$; second derivative settling characteristics.

If a signal of the form

$$E(t) = a_0 + a_1 t + \frac{1}{2} a_2 t^2$$

were to be applied suddenly to the second-derivative circuit at $t = 0$ the response would be

$$V(t) = \frac{a_0}{T^2} A_0 \left(\frac{t}{T} \right) + \frac{a_1}{T} A_1 \left(\frac{t}{T} \right) + a_2 A_2 \left(\frac{t}{T} \right)$$

where A_0 , A_1 , A_2 stand for the responses shown in Figures 2, 3, and 4, respectively, and where t is the time in seconds and T is the nominal smoothing time. The response $V(t)$ is the indicated acceleration of the target.

The sudden application of the instantaneous position and velocity components of the signal to the second-derivative circuit will give rise to some very serious consequences unless special measures are taken to mitigate them. To see this let it be assumed that $T = 20$ seconds and that the target is at such a range that $a_0 = 20,000$ yards when the signal $E(t)$ is applied to the second-derivative circuit. Each unit of A_0 in the ordinate scale of Figure 2 then represents an indicated acceleration of 50 yd per sec². Referring to Figure 2 it is clear not only that the effective settling time will be several times the smoothing time but also that the indicated acceleration will go through exceedingly large maxima.

Exceedingly large transient responses are not peculiar to second-derivative circuits. They occur also in first-derivative circuits in linear prediction, where they are due entirely to the initial position term in the signal. In all cases they are reduced to harmless proportions by special arrangements of the circuits during the operation of slewing.

13.1.3 Effect of Tracking Errors on Accuracy of Prediction

The statistical effect of tracking errors on the accuracy of prediction is most readily determined from the power spectrum of the tracking errors and the transmission function of the prediction circuit.

Table 1 gives the values of the transmission function Y_1 of the first-derivative circuit in the M9 director, referred to a nominal smoothing time of 1 second,^a and the transmission func-

tion Y_2 of the experimental second-derivative circuit design, also referred to a nominal smoothing time of 1 second. The transmission function of the linear prediction circuit with 10-second smoothing of first derivative is then

$$Y_1(p) = 1 + \frac{G_1 \cdot Y_1(10p)}{10}$$

TABLE 1*

9f	Y_1		Y_2	
	i		i	
1	0.174	0.666	-0.454	0.165
2	0.651	1.166	-1.442	1.212
3	1.312	1.358	-2.014	3.527
4	1.943	1.203	-1.069	6.688
5	2.382	0.821	2.000	9.409
6	2.599	0.364	6.575	10.115
7	2.637	-0.067	10.893	8.220
8	2.558	-0.429	13.468	4.695
9	2.416	-0.711	14.096	0.953
10	2.242	-0.920	13.401	-2.092
11	2.062	-1.070	12.064	-4.320
12	1.885	-1.172	10.530	-5.777
13	1.720	-1.238	9.027	-6.704
14	1.566	-1.279	7.652	-7.169
15	1.429	-1.299	6.438	-7.398
16	1.305	-1.304	5.382	-7.446
17	1.194	-1.299	4.471	-7.374
18	1.096	-1.286	3.683	-7.221
19	1.004	-1.268	3.015	-7.025
20	0.926	-1.247	2.436	-6.795
22	0.790	-1.198	1.509	-6.292
24	0.683	-1.145	0.818	-5.780
26	0.593	-1.091	0.301	-5.287
28	0.518	-1.040	0.088	-4.826
30	0.457	-0.991	-0.380	-4.402
32	0.407	-0.945	-0.599	-4.016
34	0.364	-0.902	-0.762	-3.666
36	0.326	-0.862	-0.881	-3.348
38	0.296	-0.825	-0.967	-3.062
40	0.266	-0.790	-1.026	-2.800

* f is in c when smoothing time $T = 1$ sec. For T -second networks, values of $9f$ are multiples of $1/9T$ c, values of Y_1 should be divided by T , and values of Y_2 should be divided by T^2 . The two networks may have different values of T .

while that of the quadratic prediction circuit with 20-second smoothing of second derivative is

$$Y_2(p) = Y_1(p) + \frac{G_2 \cdot Y_2(20p)}{400}$$

where G_1 and G_2 are determined in accordance with the discussion in Section A.10. Since

$$Y_1(p) = p(1 - 0.3724p + \dots)$$

$$Y_2(p) = p^2(1 - \dots)$$

we get

$$G_1 = t_f$$

$$G_2 = \frac{1}{2} t_f^2 + 3.724 t_f$$

^a $Y_1(p) = p \left(\frac{0.9494}{p + 1.6} - \frac{8.677}{p + 2.4} + \frac{34.74}{p + 3.6} - \frac{27.01}{p + 4.8} \right)$

Table 2 gives the values of $|Y_i(p)|^2$ and of $|Y_q(p)|^2$ for $t_f = 5, 10, 15, 20$ seconds. These are plotted in Figures 5, 6, 7, and 8.

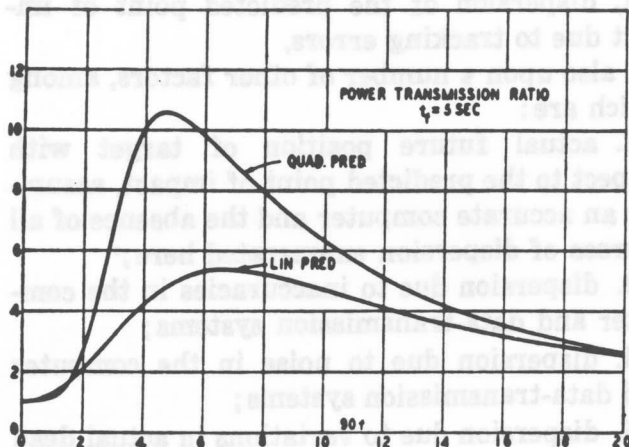


FIGURE 5. Power transmission ratio of linear and quadratic prediction circuits with 5-second prediction time.

The last column of Table 2 and Figure 9 give the power spectrum of a composite of the range and transverse errors in a typical run made with an experimental Mark VII radar. The power contained in the frequency range covered by the table accounts for 78 per cent

of the total power, or an rms error of 15.8 yards out of 17.9 yards.

The rms error of prediction is the square root of the power transmitted by the prediction circuit. This is tabulated on the last line of Table 2 and in the smaller table following.

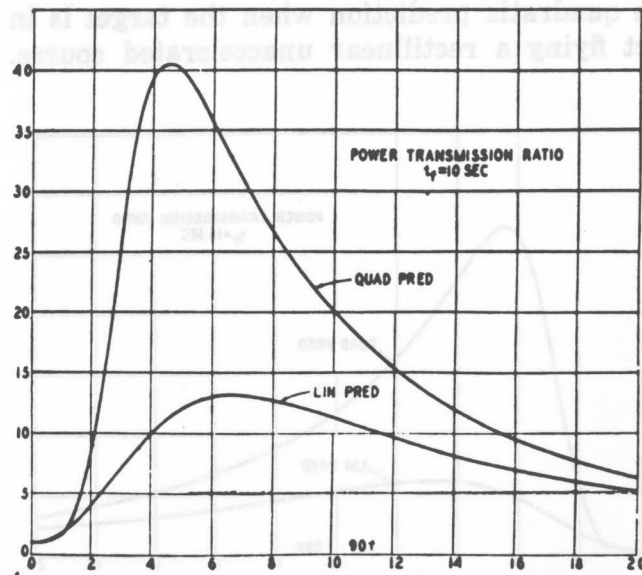


FIGURE 6. Power transmission ratio of linear and quadratic prediction circuits with 10-second prediction time.

TABLE 2

90°f	$t_f = 5$		10		15		20		P^* Mk-VII
	$ Y_i ^2$	$ Y_q ^2$	$ Y_i ^2$	$ Y_q ^2$	$ Y_i ^2$	$ Y_q ^2$	$ Y_i ^2$	$ Y_q ^2$	
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	31.4
1	1.29	1.13	1.82	1.60	2.59	2.71	3.59	4.81	33.5
2	2.10	2.76	4.08	8.90	6.97	23.16	10.74	50.35	35.7
3	3.20	6.85	7.19	26.73	12.96	72.51	20.51	159.43	19.7
4	4.2	10.0	10.1	39.5	18.6	106.1	29.76	231.3	3.6
5	5.0	10.5	12.1	39.9	22.4	104.4	35.9	223.9	2.5
6	5.3	9.8	13.1	35.6	24.3	90.6	38.9	190.6	1.2
7	5.4	8.8	13.2	30.8	24.6	76.6	39.4	158.4	1.6
8	5.2	7.9	12.8	26.6	23.8	64.7	38.2	131.8	2.1
9	5.0	7.1	12.2	23.0	22.5	55.0	36.0	110.6	1.4
10	4.7	6.3	11.4	20.0	21.0	47.0	33.5	93.5	0.7
11	4.4	5.7	10.5	17.5	19.3	40.4	30.8	79.6	0.8
12	4.1	5.1	9.7	15.3	17.7	35.0	28.3	68.2	0.8
13	3.8	4.6	8.9	13.5	16.3	30.4	25.8	58.9	0.5
14	3.6	4.2	8.2	12.1	14.9	27.1	23.6	52.0	0.3
15	3.4	3.8	7.6	10.6	13.7	23.4	21.6	44.5	0.8
16	3.2	3.5	7.0	9.5	12.6	20.6	19.8	39.0	1.1
17	3.0	3.2	6.5	8.5	11.6	18.3	18.2	34.4	0.8
18	2.8	3.0	6.0	7.7	10.7	16.3	16.8	30.4	0.4
19	2.7	2.8	5.6	7.0	9.9	14.6	15.5	27.0	0.7
20	2.5	2.6	5.3	6.3	9.2	13.1	14.4	24.1	1.0
rms error of prediction	23.9	29.5	33.9	53.4	44.5	85.4	55.4	125.0	

* P is in units of 180 yd^2 per c.

CONFIDENTIAL

Time of flight in seconds	Rms error of prediction due to tracking errors in yards	
	Linear	Quadratic
5	23.9	29.5
10	33.9	53.4
15	44.5	85.4
20	55.4	125.0

It is obviously relatively disadvantageous to use quadratic prediction when the target is in fact flying a rectilinear unaccelerated course.

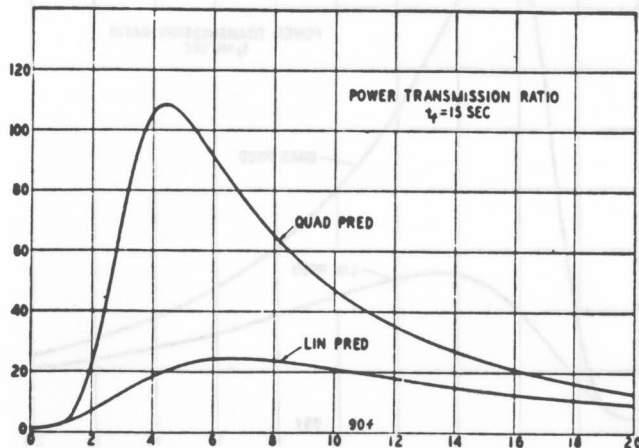


FIGURE 7. Power transmission ratio of linear and quadratic prediction circuits with 15-second prediction time.

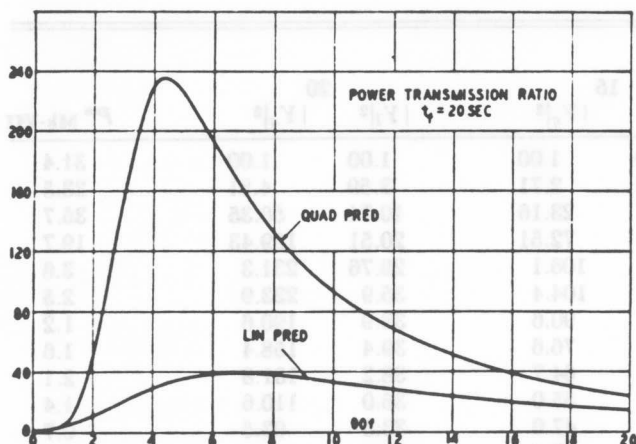


FIGURE 8. Power transmission ratio of linear and quadratic prediction circuits with 20-second prediction time.

The relative advantage of linear prediction should persist for target paths with only a slight amount of curvature, but this relative advantage should decrease as the curvature is increased. When the curvature exceeds a certain amount, the relative advantage should shift to quadratic prediction.

The determination of the minimum value of

target path curvature at which quadratic prediction becomes relatively advantageous depends not only upon:

1. dispersion of the predicted point of impact due to tracking errors, but also upon a number of other factors, among which are:

2. actual future position of target with respect to the predicted point of impact, assuming an accurate computer and the absence of all sources of dispersion enumerated here;^c

3. dispersion due to inaccuracies in the computer and data-transmission systems;

4. dispersion due to noise in the computer and data-transmission systems;

5. dispersion due to variations in actual dead time;

6. dispersion due to gun wear and to variations in powder charge, shell weight, shell shape, etc.;

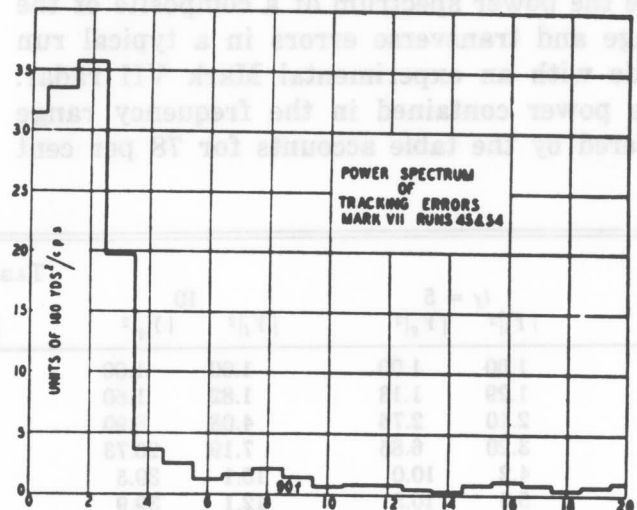


FIGURE 9. Composite power spectrum of tracking errors of experimental radar.

7. dispersion due to variations in meteorological conditions along the path of the shell;

8. dispersion due to variability of time-fuze calibration; and

9. lethal pattern of shell burst.

In a special illustrative case, a numerical analysis, including most of these factors (estimated), showed that quadratic prediction becomes relatively advantageous when the target acceleration exceeds about $0.1g$. However, this should not be taken as a general result.

^c This is considered in detail in the next section.

13.1.4 Linear and Quadratic Prediction Errors on Constant-Velocity Circular Courses

The use of a finite number of derivatives of the tracking data for purposes of prediction is itself a source of prediction errors even if there were no tracking errors. Definite evaluation of these prediction errors can be made only if the path of the target is prescribed. The simplest path which can be prescribed for this purpose is a circular one at constant velocity. Such a path is fairly realistic when considered in relation to the difficulty of maneuvering a bomber and to actual records of the paths of hostile bombers over London during World War II.

The position of a target flying in a circle at constant velocity, referred to the center of the circle, is expressed by the complex quantity $Re^{i\omega t}$ where R is the radius of the circle and ω is the angular rate. In terms of the velocity V and the transverse acceleration A , we have $R = V^2/A$ $\omega = A/V$. The predicted position is then at $RY(i\omega)e^{i\omega t}$ where $Y(i\omega)$ is the transmission function of the prediction circuit. The true future position of the target, however, is at $R \exp[i\omega(t + t_f)]$. Hence, the prediction error, referred to axes fixed on the target and oriented respectively transverse to and in the direction of the present velocity, is

$$\epsilon = R[Y(i\omega) - e^{i\omega t_f}].$$

As an illustration let us consider a case in which $V = 150$ yd per sec, $A = 5$ yd per sec² and $t_f = 10$. For the linear prediction circuit

$$Y_l(i\omega) = 1.0409 + i0.3296$$

and for the quadratic prediction circuit

$$Y_q(i\omega) = 0.9501 + i0.3610$$

while

$$e^{i\omega t_f} = 0.9450 + i0.3272.$$

Hence, when the present position of the target is at $4500 + i0$ with respect to the center of the circle, the linear predicted point is at $4684 + i1483$, the quadratic predicted point is at $4276 + i1624$ while the true future position is at $4252 + i1472$. These are shown in Figure 10. The prediction error vectors are

$$\epsilon_l = 432 + i11 \quad |\epsilon_l| = 432$$

$$\epsilon_q = 24 + i152 \quad |\epsilon_q| = 154$$

Referring to Figure 10 it may be observed that if the first-derivative component of the prediction were to be reduced by approximately 10 per cent a nearly perfect hit would be obtained. This suggests the possibility of deter-

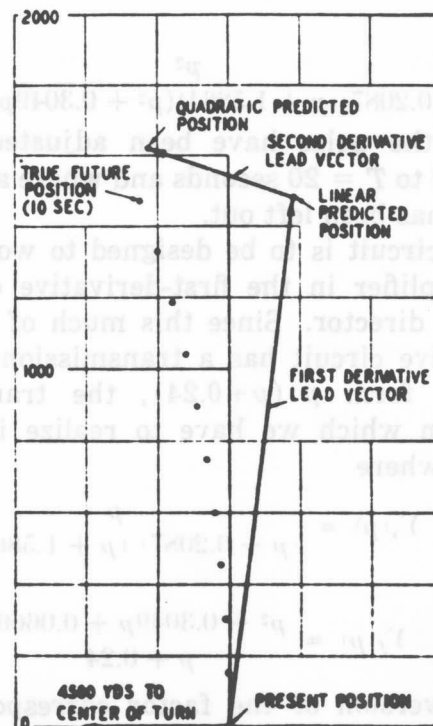


FIGURE 10. Vector diagram of linear and quadratic prediction for constant-velocity circular courses.

mining empirical functions of the time of flight for the potentiometer factors G_1 and G_2 in order to improve the probability of kill. This would involve consideration of all of the sources of dispersion enumerated in the preceding section as well as a statistical study of target paths. Such a determination has not been attempted.

13.1.5 Physical Configuration of the Second-Derivative Circuit

In this section we shall derive a physical configuration for the second-derivative circuit. In particular it illustrates the application of feedback to the realization of weighting functions or impulsive admittances involving complex exponentials in general.⁴ It should be pointed out, however, that the application of feedback to the end in view is not restricted to purely

⁴ Originally proposed by R. L. Dietzold.

electronic circuits. An application involving the use of servomechanisms will be described in Section 13.4.

The transmission function which concerns us here may be expressed in the partially factored form

$$Y(p) = \frac{p^2}{(p + 0.2087)(p + 1.5864)(p^2 + 0.3049p + 0.0666)}$$

where the poles have been adjusted to correspond to $T = 20$ seconds and where a constant factor has been left out.

The circuit is to be designed to work out of the amplifier in the first-derivative circuit of the M9 director. Since this much of the first-derivative circuit has a transmission function of the form $p(p + 0.24)$, the transmission function which we have to realize is $Y_i(p)/Y_f(p)$ where

$$Y_i(p) = \frac{p}{p + 0.2087(p + 1.5864)}$$

and

$$Y_f(p) = \frac{p^2 + 0.3049p + 0.0666}{p + 0.24}$$

The inversion of the factor corresponding to $Y_f(p)$ is in accordance with the fact that the transmission gain through a feedback amplifier is equal to the loss in the feedback network, provided the feedback is very large. To realize the transmission function $Y_i(p)/Y_f(p)$ it is therefore necessary only to realize the trans-

The input network has four elements, whereas $Y_i(p)$ has only two parameters. Hence there are two degrees of freedom in the element values of this network. One degree of freedom must be reserved for the impedance level; the other permits some latitude in the relative values of the resistances and stiffnesses.

The feedback network has four independent elements, whereas $Y_f(p)$ has three parameters. Hence there is only one degree of freedom in the element values of this network. This degree of freedom must be reserved for the impedance level.

There is, however, one degree of freedom between the impedance levels of the two networks. This follows from the fact that the transmission function of the circuit is the ratio of the transmission functions of the individual networks. The scale factor for the transmission function of the circuit is readily determined from the fact that the transmission function must be approximately pR_0C_0 at small values of p .

13.2 CIRCUIT FOR CLOSE SUPPORT PLOTTING BOARD

In this application, position data smoothing with delay correction for constant rates of change in position was required. Assuming flat random noise in position data, and, arbitrarily, 1-second smoothing time, the best transmission function for position data smoothing without delay correction is $y_0(p)$ in the notation of Section 11.3. The best transmission function for the first-derivative circuit, if it were required, is $py_1(p)$. Hence, the best transmission function for position data smoothing with full delay correction is

$$Y_I(p) = y_0(p) + \frac{1}{2} py_1(p).$$

This corresponds to the weighting function

$$\begin{aligned} W_I(t) &= w_0(t) + \frac{1}{2} \dot{w}_1(t) \\ &= 2(2 - 3t) \quad 0 < t < 1. \end{aligned}$$

The series expansion for $Y_I(p)$ is, by (15) of Chapter 11,

$$Y_I(p) = 1 - \frac{p^2}{12} + \frac{p^3}{30} - \frac{p^4}{120} + \dots$$

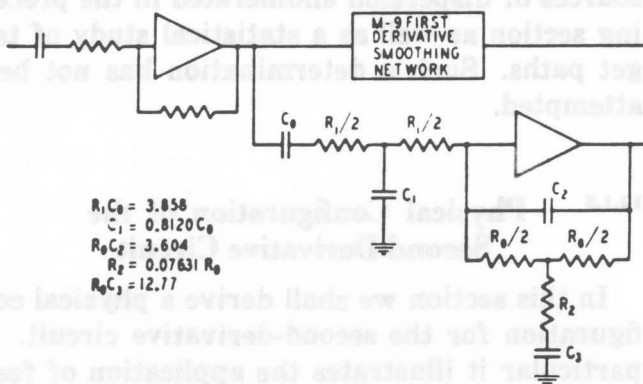


FIGURE 11. Physical configuration of quadratic prediction circuit for modified M9 AA director.

mission functions $Y_i(p)$ and $Y_f(p)$ individually. The corresponding networks are shown in Figure 11, with typical element values.

The form of the rational approximation was chosen as

$$Y(p) = \frac{1 + a_1 p}{1 + b_1 p + b_2 p^2 + b_3 p^3}$$

in order to obtain a loss characteristic which has an ultimate slope of 12 db per octave.* This requirement was also set as a precaution against noise due to granularity of the coordinate-conversion potentiometers. The coefficients are determined by

$$b_1 = a_1$$

$$b_2 - \frac{1}{12} = 0$$

$$b_3 - \frac{1}{12}b_1 + \frac{1}{30} = 0$$

$$-\frac{1}{12}b_2 + \frac{1}{30}b_1 - \frac{1}{120} = 0$$

whence

$$Y(p) = \frac{1 + \frac{11}{24}p}{1 + \frac{11}{24}p + \frac{1}{12}p^2 + \frac{7}{1440}p^3}$$

This may be expressed in the form $Y(p) = Y_i(p)/Y_f(p)$ where

$$Y_i(p) = \frac{1}{1 + 0.1053p}$$

$$Y_f(p) = \frac{1 + 0.3530p + 0.04615p^2}{1 + 0.4583p}$$

The circuit configuration is shown below in Figure 12.

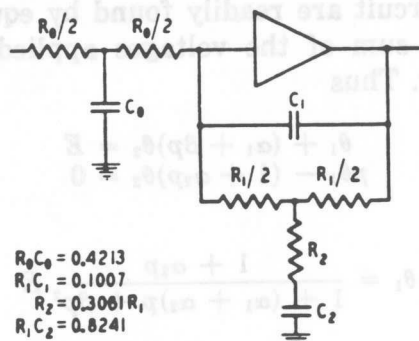


FIGURE 12. Physical configuration of data-smoothing circuit for close support plotting board.

* This design also antedated the formulation of the $n - m = r + 1$ rule given in Section 12.2 according to which we should have taken $Y_f(p) = y_1(p) + \frac{1}{2} p y_2(p)$.

13.3 CIRCUIT FOR GROUND-CONTROL BOMBING COMPUTER

In this application, rate smoothing as well as position smoothing was required. In addition, delay correction in position, for constant rate of change, was to be available but optional, and the loss characteristic was to have an ultimate slope of 12 db per octave, or more.

In accordance with the $n - m = r + 1$ rule, the best transmission function for position data is $y_1(p)$, whereas that for rate is $p y_2(p)$. A number of designs were made on this basis. However, from the point of view of network economy, they were inferior to a design based on $y_2(p)$ for position data. The use of $y_2(p)$ for position data is not consistent, theoretically, with the use of $p y_2(p)$ for rate, but the practical advantage outweighs the theoretical disadvantage.

The rational approximation used for $y_2(p)$

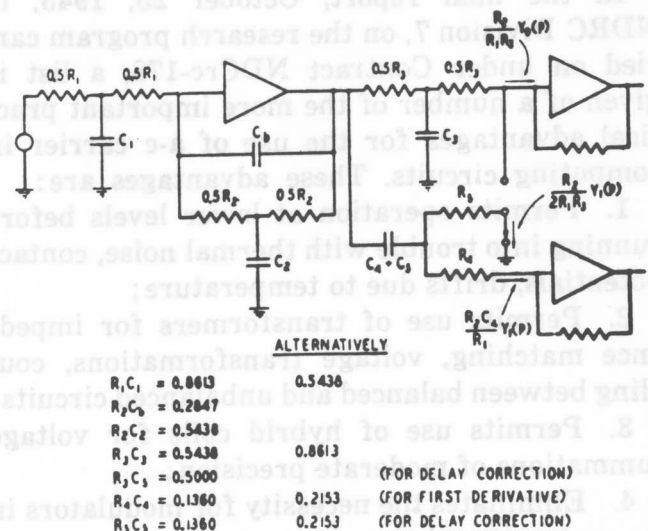


FIGURE 13. Physical configuration of linear prediction circuit for ground-control bombing computer.

is the one given in (6), Section 12.2. It may be expressed as

$$y_2(p) = \frac{Y_{i1}(p) \cdot Y_{i2}(p)}{Y_f(p)}$$

where

$$Y_{i1}(p) = \frac{1}{1 + 0.2153p}$$

$$Y_f(p) = \frac{1 + 0.2847p + 0.03870p^2}{1 + 0.1359p}$$

$$Y_{i2}(p) = \frac{1}{1 + 0.1359p}$$

It may be noted that a redundant factor has been introduced, viz., $1 + 0.1359p$, in order to secure a physically realizable $Y_i(p)$. The coefficient was chosen so that a resistance would not be required in the shunt branch of the feedback network. Referring to the circuit configuration in Figure 13, the transmission function of the input network is $Y_{i1}(p)$, that of the feedback network is $Y_f(p)$, and that of the output network at the top is $Y_{i2}(p)$.

The output impedance of the amplifier is reduced nearly to zero by virtue of shunt feedback.¹⁵⁾ Hence, the rate circuit, as shown in Figure 13, may be derived from the amplifier output through a simple additional network whose transmission function is $pY_{i2}(p)$. Two rate outputs are provided so that the delay introduced in position may be corrected optionally without disturbing scale factors.

13.4 CIRCUIT USING SERVOMECHANISMS

In the final report, October 25, 1945, to NDRC Division 7, on the research program carried on under Contract NDCrc-178, a list is given of a number of the more important practical advantages for the use of a-c carrier in computing circuits. These advantages are:

1. Permits operation at lower levels before running into trouble with thermal noise, contact potentials, drifts due to temperature;
2. Permits use of transformers for impedance matching, voltage transformations, coupling between balanced and unbalanced circuits;
3. Permits use of hybrid coils for voltage summations of moderate precision;
4. Eliminates the necessity for modulators in servo circuits using a-c motors;
5. Permits reduction in total power consumption, rectified power for amplifiers, and voltage regulation.

However, the techniques of differentiation and of data smoothing with fixed networks in computing circuits which use d-c carrier, are not applicable to computing circuits which use a-c carrier.

The circuit described here is an example of one of the techniques used in the T15-E1 experimental curved flight director.⁶ In Figure 14 servo motors⁷ are indicated by M , and genera-

tors by G . The motors are two-phase induction motors with one phase winding of each energized directly by the carrier source at constant amplitude. The generators are essentially two-phase induction motors also with one phase winding of each energized directly by the carrier source at constant amplitude. They deliver, at

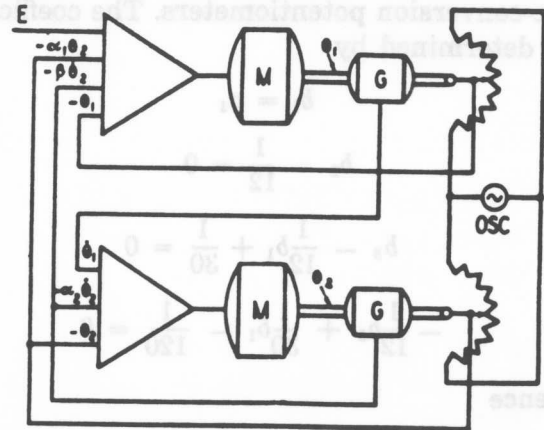


FIGURE 14. Electromechanical linear prediction circuit.

the other phase windings, carrier voltage at amplitudes proportional to the angular velocities $\dot{\theta}_1$ and $\dot{\theta}_2$ of the shafts. The potentiometers are energized by the carrier source at constant amplitude. They deliver carrier voltage at amplitudes proportional to the angular positions θ_1 and θ_2 of the shafts from some reference positions. The position data are represented by the modulation amplitude E .

With amplifiers of sufficiently large voltage gain and power capacity, and motors of sufficiently large torque, the operational equations of the circuit are readily found by equating to zero the sum of the voltages applied to each amplifier. Thus

$$\begin{aligned}\theta_1 + (\alpha_1 + \beta p)\theta_2 &= E \\ p\theta_1 - (1 + \alpha_2 p)\theta_2 &= 0\end{aligned}$$

whence

$$\theta_1 = \frac{1 + \alpha_2 p}{1 + (\alpha_1 + \alpha_2)p + \beta p^2} E$$

$$\theta_2 = \frac{p}{1 + (\alpha_1 + \alpha_2)p + \beta p^2} E$$

The angular position θ_1 therefore represents the smoothed position data while the angular position θ_2 represents the smoothed rate.

⁷ The technique of using servo motors for smoothing, as described above, is due chiefly to E. L. Norton.

VARIABLE AND NONLINEAR CIRCUITS

THE PAST DISCUSSION has been more or less clearly directed at predictor systems having certain well-defined properties. For example, it has been tacitly assumed that the first part of the prediction system will consist of geometrical manipulations transforming the raw input data into other quantities, such as the components of velocity in Cartesian or intrinsic coordinates, which we have some physical reason to believe should be approximately constant for extended periods.^a These quantities, then, are isolated explicitly in the circuit and are the actual effective inputs of the data-smoothing networks. The data-smoothing networks themselves are, of course, definitely assumed to be linear and invariable.

This is obviously a straightforward attack but it does not necessarily exhaust all possibilities. For example, advantages may be gained by using data-smoothing networks which are nonlinear or which vary with time or target position. It may also be possible to smooth the input data according to some geometric assumption, such as straight line flight, without the necessity of isolating geometrical parameters explicitly.

This chapter attempts to illustrate these possibilities by some rather scattered examples. Data-smoothing networks which vary with time seem to give improved performance over fixed networks, and have been studied with some care. Several examples are given at the end of the chapter. None of the other lines, however, has been explored at all thoroughly. The examples of data-smoothing networks variable with time are, in a sense, illustrations of nonlinearity also, since they all operate on the assumption that the cycle of the network's variation with time begins anew at each marked change in course. Since a change in course is exactly like a tracking error, except that it is much larger, this resetting requires a nonlinear control circuit which will respond to large amplitude effects but not to small ones.

^a This is true ideally even in the Wiener system since Wiener assumes that transformations will be made to some suitable coordinate system, preferably the intrinsic, before the statistical prediction method is applied.

This, however, is evidently a very mild sort of nonlinearity. More thoroughgoing nonlinearities have not been studied. There seems to be no *a priori* reason for supposing that they would appreciably improve the performance of data-smoothing networks.

The first part of the chapter gives examples of data-smoothing schemes which do not require the isolation of geometrical parameters. They are based on degenerative feedback circuits which satisfy the requisite formal relations but which might, in some cases, be unstable in practice. This portion of the material is included primarily for its possible suggestive value rather than for its concrete practical usefulness.

14.1 THE PROTOTYPE FEEDBACK CIRCUIT

The diversity of particular circuits can be given a certain unity by regarding them all as modifications of the feedback smoothing circuit shown originally in Figure 2 of Chapter 10. In accordance with the discussion of that figure it will be convenient to suppose that the resistive feedback path is introduced to limit the gain of the amplifier proper, so that the structure reduces to an amplifier with high but finite gain and a pure capacity feedback. The circuit has a net loop gain, and is consequently degenerative, at any moderately high frequency. For our present purposes, it is convenient to recall the general property of degenerative feedback amplifiers, that they tend to suppress any given frequency by the amount of the degenerative feedback for that frequency. This suppression obtains not only at the amplifier output but at many other points in the circuit as well. For example, it holds at the amplifier input if we combine the original applied voltage with the voltage contributed by the feedback^b circuit.^{15g} Thus, except for the absolute

^b This follows immediately from the fact that, since the characteristics of the amplifier proper are not changed by the addition of the feedback path, the output voltage is always a fixed multiple of the net input voltage.

signal level, it is not necessary to transmit through the amplifier of Figure 2 of Chapter 10 in order to produce the smoothing effect. It would be sufficient to hang the input circuit of the amplifier, as a two-terminal impedance, across the circuit.

14.2 SIMULTANEOUS SMOOTHING IN THREE COORDINATES

The property of degenerative feedback circuits which has just been described is conveniently illustrated by a three-dimensional extension of the original smoothing circuit of Figure 2 of Chapter 10. The three-dimensional circuit is shown in Figure 1. The three input voltages are the quantities D , $D\dot{E}$, and $D\dot{A} \cos E$

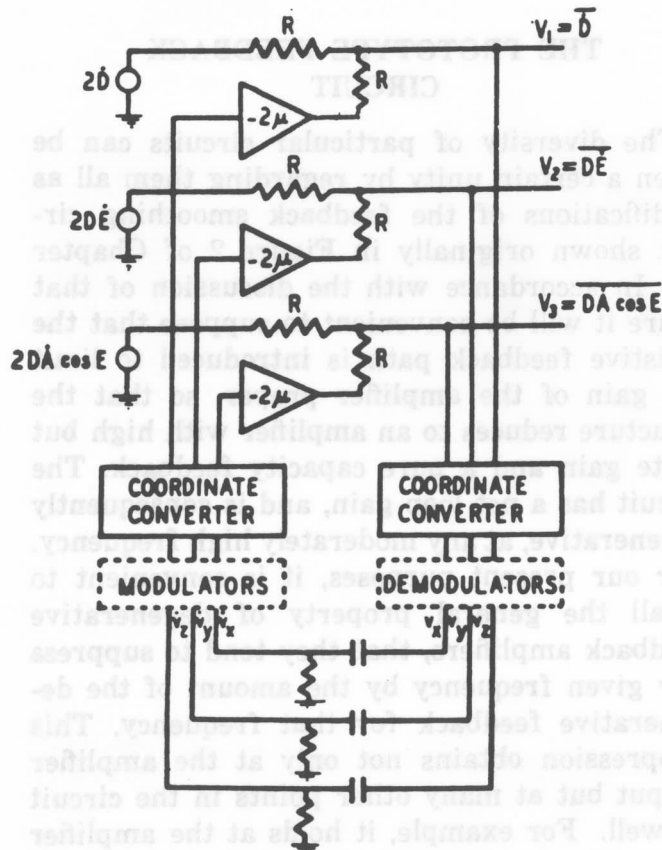


FIGURE 1. Feedback smoothing in three coordinates.

E , where D , E , and A are, respectively, slant range, elevation, and azimuth. The three voltages will be recognized as the three components of the target motion in a tilted and rotating rectangular coordinate system. One axis of the tilted system is directed along the instan-

aneous line of sight to the target and the other two are perpendicular to this one in the vertical and horizontal planes respectively.* It is assumed that these input rates represent target motion in a straight line, plus the usual tracking errors. The object of the smoothing system is to provide shunt impedances which will tend to suppress the tracking errors by feedback action, according to the principles described in the preceding section, without disturbing the portions of the input voltages corresponding to the assumed straight line path.

We can simplify the analysis by restricting our attention to the special case of two-dimensional motion which occurs when the target course lies in a vertical plane passing directly through the antiaircraft position. This is illustrated in Figure 2. In this case the component $D\dot{A} \cos E$ is evidently zero. If we represent the voltage at the other two terminals, including both the original applied voltages and the voltages fed back through the circuit, by V_1 and V_2 , the voltages coming out of the coordinate converter on the right-hand side in Figure 2 are

$$\begin{aligned} v_s &= V_1 \cos E - V_2 \sin E \\ v_y &= V_2 \cos E + V_1 \sin E \end{aligned} \quad (1)$$

These voltages are differentiated, passed through a second coordinate converter, and fed back so that the output voltages must satisfy the equations

$$\begin{aligned} V_1 &= \dot{D} - \mu(\dot{v}_s \cos E + \dot{v}_y \sin E) \\ V_2 &= D\dot{E} - \mu(\dot{v}_y \cos E - \dot{v}_s \sin E) \end{aligned} \quad (2)$$

In order to exhibit the smoothing action of the circuit let us denote the observed velocity components, referred to the upright and fixed

* This is the coordinate system which was used in the experimental T15 director. A complete prediction circuit can be obtained by using the three voltages described here as inputs to the lead servos in the T15 system. In the actual T15 system, rates in the tilted and rotating coordinate system were obtained by the so-called "memory point" method. The voltages \dot{D} , $D\dot{E}$, etc., required with the present method, might be obtained with the help of tachometers attached to the tracking shafts to measure the instantaneous values of \dot{D} , \dot{E} , and \dot{A} . An equivalent to the variable smoothing of the memory point method can be obtained by making the gains in the feedback paths in Figure 1 variable according to the principles described in a later section.

rectangular coordinate system, by u_x and u_y , so that

$$\begin{aligned} u_x &= \dot{D} \cos E - D\dot{E} \sin E \\ u_y &= D\dot{E} \cos E + \dot{D} \sin E. \end{aligned} \quad (3)$$

Substituting (2) and (3) into (1), we get

$$v_x = u_x - \mu \dot{v}_x$$

$$v_y = u_y - \mu \dot{v}_y$$

or

$$\mu \dot{v}_x + v_x = u_x$$

$$\mu \dot{v}_y + v_y = u_y.$$

These show clearly that v_x and v_y are smoothed values of u_x and u_y , respectively. If μ is constant the smoothing is of fixed exponential type. If μ is proportional to the time up to some maximum value, the smoothing is of the variable type described in Sections 14.6 and 14.7.

To complete the discussion of the circuit we observe that by (1)

$$V_1 = v_x \cos E + v_y \sin E$$

$$V_2 = v_y \cos E - v_x \sin E.$$

These show that V_1 and V_2 are the smoothed rate components referred to the tilted and rotating rectangular coordinate system. The fact that the orientation of this coordinate system, which depends upon the observed angular height E , is not smoothed makes no difference to the computation of the leads because this computation is made instantaneously in the same coordinate system to which the smoothed rate components are instantaneously referred.

The analysis in the general case including all three coordinates is of the same nature. Since the rate components in fixed rectangular coordinates appear in the middle of the feedback path, it is perhaps not fair to regard the circuit as an illustration of a data-smoothing device which does not rely upon the explicit isolation of the geometrical parameters of the assumed target path. It should be pointed out, however, that in comparison with a straightforward geometrical solution in which velocity components in fixed coordinates are first isolated explicitly, then smoothed, and then used to form the basis of prediction, the circuit in Figure 1 has the advantage that most of the components can be built with very low precision. What is transmitted around the feedback loop is essen-

tially the tracking errors only. Since tracking errors are always small, very high percentage errors in the system can be tolerated.⁴

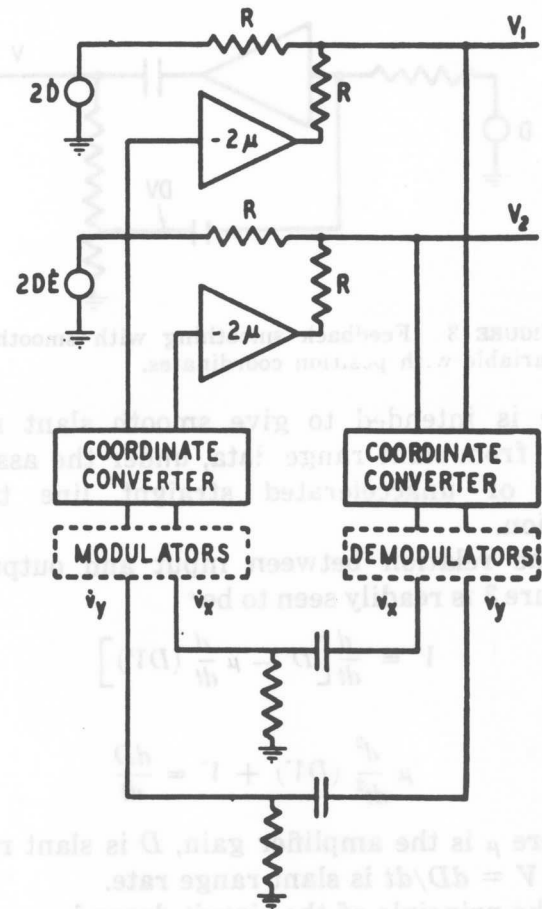


FIGURE 2. Feedback smoothing in two coordinates.

14.3 SMOOTHING NETWORKS VARIABLE WITH TARGET POSITION

It was mentioned earlier that changing the data-smoothing network with the target coordinates represented one way in which the results obtained from fixed networks could be

⁴ An exception to this statement must be made for errors in the coordinate converters which fluctuate rapidly with target position.

generalized. In a sense, the coordinate conversions of Figure 1 are illustrations of these possibilities. A better illustration, however, is provided by the circuit of Figure 3. The struc-

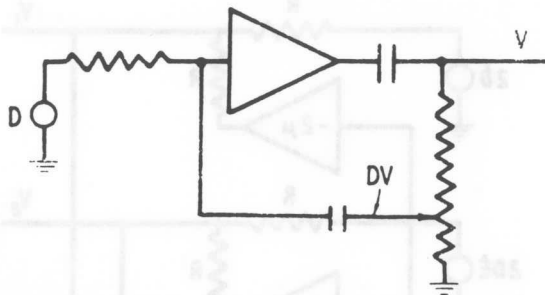


FIGURE 3 Feedback smoothing with smoothing variable with position coordinates.

ture is intended to give smooth slant range rate from slant range data, under the assumption of unaccelerated straight line target motion.

The relation between input and output in Figure 3 is readily seen to be^e

$$V = \frac{d}{dt} \left[D - \mu \frac{d}{dt} (DV) \right]$$

or

$$\mu \frac{d^2}{dt^2} (DV) + V = \frac{dD}{dt} \quad (4)$$

where μ is the amplifier gain, D is slant range, and $V = dD/dt$ is slant range rate.

The principle of the circuit depends upon the fact that under the assumed target motion the square of the slant range, D^2 , should be a quadratic function of time, so that $[D(dD/dt)]$ should be a linear function of time and $(d/dt)[D(dD/dt)]$ should be a constant. This last is the quantity which is fed back in Figure 3. If it actually is a constant, it has no further influence on the calculation, since the forward circuit includes a differentiator, and the operation of the circuit is the same as though no feedback term were present. This can be verified by setting $D = D_0 = \sqrt{a + 2bt + ct^2}$, corresponding to ideal straight line flight, in equation (4). It is readily seen that the equation is satisfied by

$$V = V_0 = \frac{b + ct}{\sqrt{a + 2bt + ct^2}} = \frac{dD_0}{dt},$$

the first or feedback term being zero.

^e The condensers in Figure 3 symbolize differentiation.

If D does not correspond exactly to straight line flight, either because of tracking errors or actual target maneuvers, on the other hand, the feedback voltage is no longer constant. In this case transmission around the loop can exist and the degenerative feedback action produces smoothing in both the input and the output voltage. In calculating the exact effect we must take account of the fact that the feedback voltage depends upon the D potentiometer in the feedback circuit as well as upon the output voltage V . Since the D potentiometer setting must include the errors in the input data, this means that the output voltage is not perfectly smoothed, even with unlimited gain around the loop. The percentage error in the output rate tends in the limit to approximate the percentage error in D itself. For practical purposes, however, this is a very satisfactory result, since in the absence of smoothing percentage errors in rates are usually many times those of the corresponding coordinates.

It is apparent that it should be possible to construct many circuits of this general type from the differential equations of the trajectory. A second example is furnished by Figure 4. The operation of the circuit is essentially

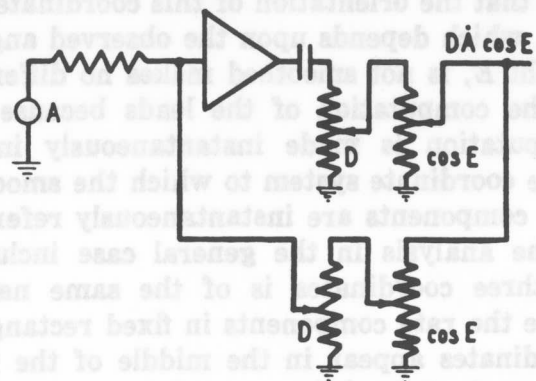


FIGURE 4. Another example of feedback smoothing with smoothing variable with position coordinates.

similar to that of Figure 3. It depends upon the fact that in unaccelerated straight line motion the quantity $D^2 \dot{A} \cos^2 E$ is a constant. Instead of multiplying by D^2 and $\cos^2 E$ at a single point in the feedback loop, however, separate multiplications by D and $\cos E$ are introduced in the forward and feedback circuits. This permits the output to appear as a smoothed value of the quantity $D \dot{A} \cos E$,

which will be recalled as one of the primary quantities in the circuit of Figure 1.

14.4 NETWORKS VARIABLE WITH TIME

In addition to making the parameters of the data-smoothing network vary as functions of the coordinates of target position we may also make them variable as functions of time. The advantage of variation with time can be understood by going back to the discussion of the analytic arc assumption and its consequences for fixed data-smoothing networks, as given in Chapters 9, 10, and 11. It will be recalled that for any given settling time there was an optimum choice of the network's weighting function. The choice of the settling time itself, however, was always a compromise. On the one hand, making the settling time too short led to too little smoothing, so that the dispersion in the resulting fire became excessive. On the other hand, too long a settling time meant that data from previous unrelated segments were retained in the smoothing circuit during too large a proportion of an average individual segment of the target path, leaving too small a residue of the average segment as useful firing time.

It is evident that it is theoretically possible to escape the consequences of this compromise by resorting to variable structures. We need merely assume that the network always has a weighting function appropriate for a settling time equal to the time since the last change in course. This would give a small amount of smoothing shortly after a change in course, with more smoothing and consequently greater accuracy later on. No firing time, however, is sacrificed waiting for the network to settle.

In order to exploit these possibilities we must, of course, be able to design networks to give at least approximately the right sequence of weighting function. It is also necessary to provide some sort of auxiliary controlling mechanism which will sense changes in target course and return the variable circuits in the smoothing network proper to their initial positions. These are both difficult problems which have been incompletely explored. Some elementary solutions, based principally upon modifications of the degenerative feedback smoothing

circuit of Figure 2, of Chapter 10, are, however, given later in the chapter. As a preliminary, the next section gives a formal extension of the general polynomial expansion method of Chapter 11 to the variable case.

14.5 GENERAL POLYNOMIAL SOLUTION FOR VARIABLE NETWORKS

The extension of the general method of Chapter 11 to the variable case requires two modifications.

1. The lower limit of the integral to be minimized is now taken as zero, in anticipation of the possibility of discriminating between relevant and irrelevant data on the basis of time of arrival.

2. The weighting function may now depend more generally upon the variable of integration and the upper limit of integration.

With these modifications there is no longer any advantage in conducting the analysis in terms of the age variable τ . To deal directly with the minimization of the integral

$$\int_0^t [\bar{E}(\lambda) - E(\lambda)]^2 W_0(t, \lambda) d\lambda, \quad (5)$$

let

$$\bar{E}(\lambda) = V_0 + V_1 \cdot G_1(t, \lambda) + \cdots + V_n \cdot G_n(t, \lambda), \quad (6)$$

Where $G_m(t, \lambda)$ is an m th degree polynomial in λ . Also, let

$$\begin{aligned} \int_0^t W_0(t, \lambda) d\lambda &= 1 \\ \int_0^t G_l(t, \lambda) \cdot G_m(t, \lambda) \cdot W_0(t, \lambda) d\lambda &= 0 \quad \text{if } l \neq m \\ &= \frac{1}{k_m} \quad \text{if } l = m \end{aligned}$$

($G_0 = 1, k_0 = 1$).

Then (5) is a minimum with respect to the V_m 's in (6) if

$$V_m(t) = \int_0^t E(\lambda) \cdot W_m(t, \lambda) d\lambda \quad (7)$$

where

$$W_m(t, \lambda) = k_m G_m(t, \lambda) \cdot W_0(t, \lambda). \quad (8)$$

The possibility of physically realizing the $V_m(t)$ depends upon the possibility of realizing networks with impulsive admittances $W_m(t, \lambda)$ in the sense that $W_m(t, \lambda)$ is the response of a

network, at time t , to a unit impulse applied at time λ , where $0 \leq \lambda \leq t$. Taking this possibility for granted, the predicted value $\bar{E}(t + t_f)$ is, according to (6), a variable linear combination of the $V_m(t)$, viz.,

$$\bar{E}(t + t_f) = V_0(t) + G_1(t, t + t_f) \cdot V_1(t) + \dots + G_n(t, t + t_f) \cdot V_n(t). \quad (9)$$

It is clear that all of the $W_m(t, \lambda)$ as well as all of the $G_m(t, \lambda)$ for $m = 1, 2, \dots$ are determined by $W_0(t, \lambda)$. The latter is determined as the best weighting function for position data smoothing, depending upon the characteristics of the noise associated with the position data. The general methods of determining the best weighting function with fixed smoothing time, described in Chapter 10, may be used to determine the best weighting function with variable smoothing time.

Under the assumption that the spectrum of the noise associated with the signal $S(t)$ has a uniform slope of $6k$ db per octave, we may take over from Section 11.3 the result that the best weighting function is

$$w_k(t, \lambda) = \frac{(2k+1)!}{(k!)^2 t} \left[\frac{\lambda}{t} \left(1 - \frac{\lambda}{t} \right) \right]^k \quad (10)$$

$$0 \leq \lambda \leq t.$$

The response of the network is then

$$V(t) = \int_0^t S(\lambda) \cdot w_k(t, \lambda) d\lambda. \quad (11)$$

14.6 SPECIAL CASES

It will be illuminating to consider a few special cases of (11).

For $k = 0$, we have

$$V(t) = \frac{1}{t} \int_0^t S(\lambda) d\lambda. \quad (12)$$

Multiplying through by t and differentiating we get

$$t\dot{V}(t) + V(t) = S(t). \quad (13)$$

This suggests the circuit shown in Figure 5.¹

For $k = 1$, we have

$$V(t) = \frac{6}{t^3} \int_0^t S(\lambda) \cdot \lambda(t - \lambda) d\lambda.$$

¹ This circuit is due to S. Darlington.

Multiplying through by t^3 and differentiating twice we get

$$\frac{1}{6} t^2 \ddot{V} + t\dot{V} + V = S$$

which may be written in the form

$$\left(\frac{t}{2} \frac{d}{dt} + 1 \right) \left(\frac{t}{3} \frac{d}{dt} + 1 \right) \cdot V = S.$$

This suggests the network shown in Figure 6.²

14.7 NETWORKS WITH A LIMITED RANGE OF VARIATION

By generalizing the above results in various ways a large number of other examples of variable smoothing networks can be constructed. Since unlimited variation in the smoothing time is not practically possible, or perhaps even tactically optimal, however, it is desirable in discussing any further examples to include also the possibility that the range of variation in the network may be restricted. For any positive integral value of k in (11) the differential equation for $V(t)$ is of the type which may be reduced by the transformation $t = e^z$ to a linear differential equation with constant coefficients.³ In general, this facilitates the determination of what happens to the weighting function $w_k(t, \lambda)$ when $t > T$ if the variability of the network is stopped at time T . In the case of the first-order equation (13), however, it is just as easy to deal directly in terms of the natural time.

A more general form for (13), which readily yields the effects of a sudden or gradual stoppage of the variability of the network, is

$$\frac{\phi(t)}{\phi(t)} \dot{V}(t) + V(t) = S(t). \quad (14)$$

This corresponds to the response

$$V(t) = \frac{1}{\phi(t)} \int_0^t S(\lambda) \cdot \phi(\lambda) d\lambda$$

whence the weighting function is

$$w(t, \lambda) = \frac{\phi(\lambda)}{\phi(t)}. \quad (15)$$

² Due to B. T. Weber.

³ See Section A.11 for a more general transformation.

The general relation (14) may be realized with the network of Figure 5, by varying the resistance in accordance with

$$R = \frac{1}{C} \frac{\phi(t)}{\dot{\phi}(t)} \quad t > 0.$$

However, a more practical circuit results from the introduction of variable potentiometers¹ in both the capacity and resistance paths of the

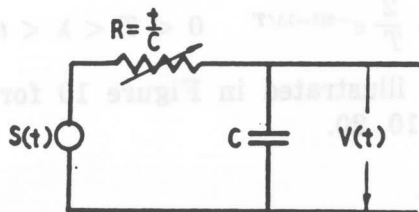


FIGURE 5. Time-variable smoothing circuit giving uniform weighting function.

original feedback smoothing circuit of Figure 2, Chapter 10. This is shown in Figure 7.¹ It may be noted that the feedback circuit is also applicable to the two cases discussed in the preceding section. It has the advantage for these applications that it does not require the zero-impedance generators and infinite-impedance loads of Figures 5 and 6.

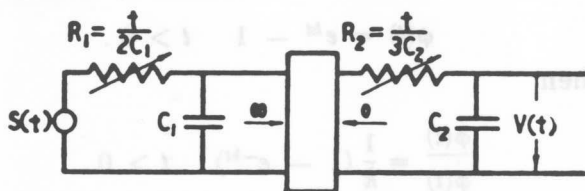


FIGURE 6. Time-variable smoothing circuit giving parabolic weighting function.

As an example of (14) we may take

$$\begin{aligned} \phi(t) &= t \quad 0 < t < T \\ &= Te^{(t-T)/T} \quad t > T. \end{aligned}$$

Then

$$\begin{aligned} \frac{\phi(t)}{\dot{\phi}(t)} &= t \quad 0 < t < T \\ &= T \quad t > T. \end{aligned}$$

Hence, in Figure 7, if $RC = T$

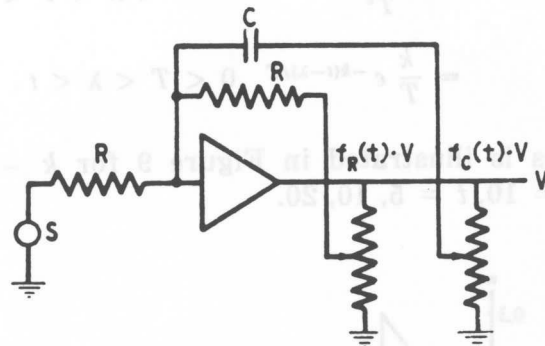
$$\begin{aligned} f_C(t) &= \frac{t}{T} \quad f_R(t) = 0 \quad 0 < t < T \\ &= 1 \quad = 1 \quad t > T. \end{aligned}$$

¹ In some cases a variable potentiometer may turn out to be a switch.

¹ This circuit is due to S. Darlington.

This example obviously calls for a linear potentiometer in the condenser path and a switch in the resistance path. The weighting function obtained is, by (15),

$$\begin{aligned} w(t, \lambda) &= \frac{1}{t} \quad 0 < \lambda < t < T \\ &= \frac{1}{T} e^{-(t-T)/T} \quad 0 < \lambda < T < t \\ &= \frac{1}{T} e^{-(t-\lambda)/T} \quad 0 < T < \lambda < t \end{aligned}$$



$$f_C(t) = \frac{1}{RC} \frac{\phi(t)}{\dot{\phi}(t)} \quad f_R(t) = 1 - \frac{d}{dt} \left[\frac{\phi(t)}{\dot{\phi}(t)} \right]$$

FIGURE 7. Limited range time-variable feedback smoothing circuit.

This is illustrated in Figure 8 for $T = 10$, $t = 5$, 10, 20.

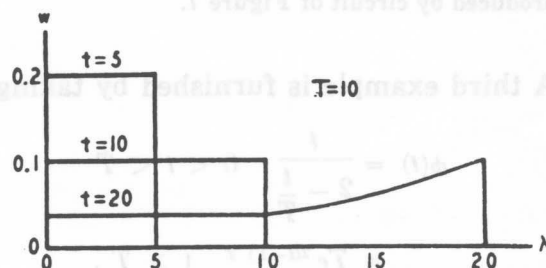


FIGURE 8. First example of weighting function produced by circuit of Figure 7.

A second example is furnished by taking

$$\begin{aligned} \phi(t) &= t^k \quad 0 < t < T \\ &= T^k e^{k(t-T)/T} \quad t > T. \end{aligned}$$

Then

$$\begin{aligned} \frac{\phi(t)}{\dot{\phi}(t)} &= \frac{t}{k} \quad 0 < t < T \\ &= \frac{T}{k} \quad t > T. \end{aligned}$$

Hence in Figure 7, if $RC = T/k$,

$$f_c(t) = \frac{t}{T} \quad f_R(t) = 1 - \frac{t}{T} \quad 0 < t < T$$

$$= 1 \quad = 1 \quad t > T.$$

The first example is a special case of this one. The weighting function obtained is, by (15),

$$w(t, \lambda) = \frac{k\lambda^{k-1}}{t^k} \quad 0 < \lambda < t < T$$

$$= \frac{k\lambda^{k-1}}{T^k} e^{-k(t-T)/T} \quad 0 < \lambda < T < t$$

$$= \frac{k}{T} e^{-k(t-\lambda)/T} \quad 0 < T < \lambda < t.$$

This is illustrated in Figure 9 for $k = 3/2$, $T = 10$, $t = 5, 10, 20$.

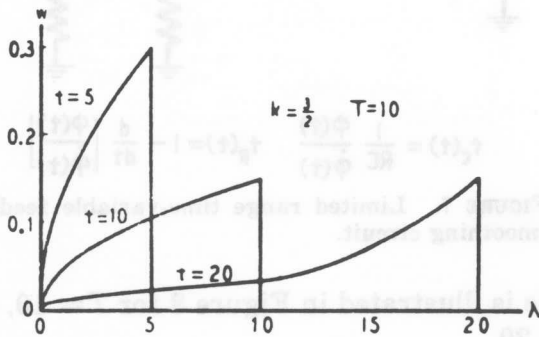


FIGURE 9. Second example of weighting function produced by circuit of Figure 7.

A third example is furnished by taking

$$\phi(t) = \frac{t}{2 - \frac{t}{T}} \quad 0 < t < T$$

$$= T e^{2(t-T)/T} \quad t > T.$$

Then

$$\frac{\phi(t)}{\phi(t)} = t \left(1 - \frac{t}{2T} \right) \quad 0 < t < T$$

$$= \frac{T}{2} \quad t > T.$$

Hence, in Figure 7, if $RC = T/2$,

$$f_c(t) = \frac{2t}{T} \left(1 - \frac{t}{2T} \right) \quad f_R(t) = \frac{t}{T} \quad 0 < t < T$$

$$= 1 \quad = 1 \quad t > T.$$

The weighting function obtained is, by (15),

$$w(t, \lambda) = \frac{1 - \frac{t}{2T}}{t \left(1 - \frac{\lambda}{2T} \right)^2} \quad 0 < \lambda < t < T$$

$$= \frac{1}{2T \left(1 - \frac{\lambda}{2T} \right)^2} e^{-2(t-T)/T} \quad 0 < \lambda < T < t$$

$$= \frac{2}{T} e^{-2(t-\lambda)/T} \quad 0 < T < \lambda < t.$$

This is illustrated in Figure 10 for $T = 10$, $t = 5, 10, 20$.

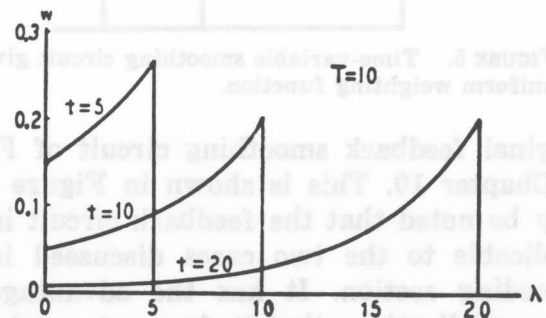


FIGURE 10. Third example of weighting function produced by circuit of Figure 7.

A fourth example is furnished by taking

$$\phi(t) = e^{kt} - 1 \quad t > 0.$$

Then

$$\frac{\phi(t)}{\phi(t)} = \frac{1}{k} (1 - e^{-kt}) \quad t > 0.$$

Hence, in Figure 7, if $RC = 1/k$,

$$f_c(t) = f_R(t) = 1 - e^{-kt} \quad t > 0.$$

The weighting function obtained is, by (15),

$$w(t, \lambda) = \frac{k}{1 - e^{-kt}} e^{-k(t-\lambda)} \quad 0 < \lambda < t.$$

For any value of t this weighting function is exponential in λ .

Because there has been no demand for variable networks in the field of communications, the technique of designing practical variable networks is in a very rudimentary stage compared to that of designing fixed networks. In the remainder of this chapter we shall describe

some of the circuits which have been developed for specific practical applications.

A memory point method of obtaining smoothed rates, based upon (12), is illustrated below. If $S(t)$, the quantity to be smoothed, represents the time derivative $\dot{E}(t)$ of the position data $E(t)$, then the average rate is given by

$$\bar{S}(t) = \frac{E(t) - E(0)}{t} \quad (16)$$

Under the assumption that the position data, aside from tracking errors, is a linear function of time, the average rate is also the smoothed rate. If the position data is represented by the angular displacement of a shaft in the computer, the quantity $E(0)$ is readily fixed by providing a second shaft which is coupled to the first shaft until $t = 0$ when the coupling is broken. Potentiometers mounted on the shafts are energized by a voltage varying as a function of time in the manner indicated in Figure 11. The manner in which the smoothed rate is obtained is clear.

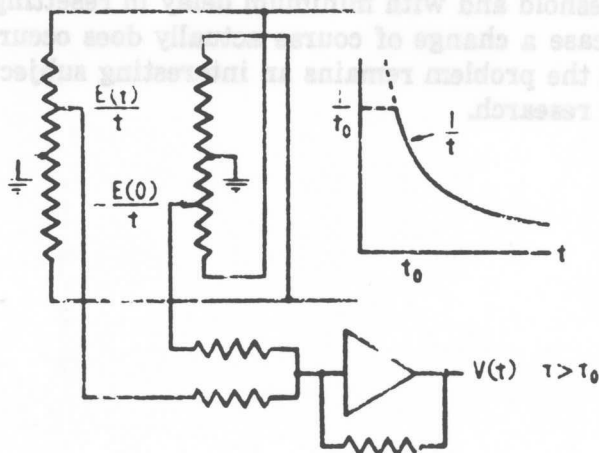


FIGURE 11. Memory point method of obtaining smoothed rate.

The memory point method of obtaining smoothed rates is used in the T15 antiaircraft director.⁴ In this application, however, it is somewhat more complicated than in the simple illustration described above. This is due to the fact that the position data and the memory point are in the polar coordinate system, whereas the rate components are referred to a tilted and rotating rectangular coordinate system which is determined by the instantaneous line of sight.

Figure 12, shows a way of securing variable smoothing in a purely electrical circuit.^{*} Except for the fact that the division of the current through the condensers is varied discontinu-

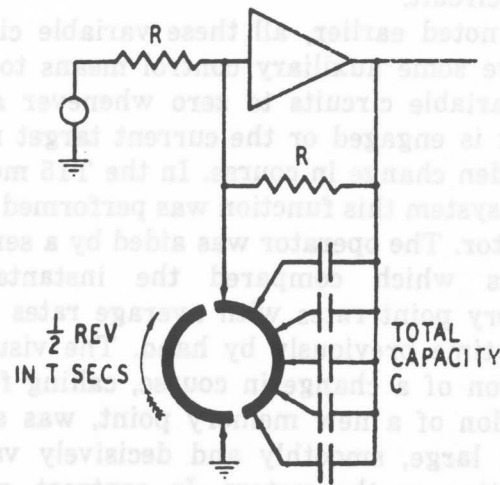


FIGURE 12. Specific limited range time-variable feedback smoothing circuit.

ously instead of continuously, this circuit corresponds to the first or the second example discussed in Section 14.7.

Figure 13 shows the variable smoothing circuit¹ for smoothing first derivatives in the M9A1-E1 antiaircraft director.⁸ This circuit

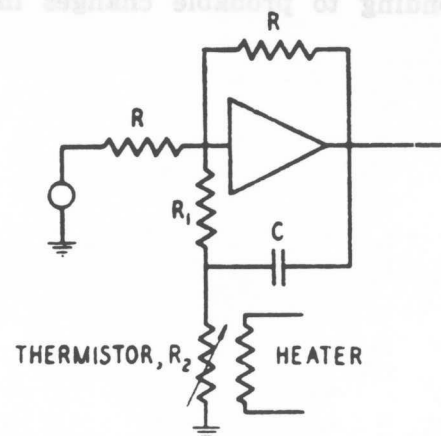


FIGURE 13. Another specific limited range time-variable feedback smoothing circuit.

corresponds approximately to the second example of the differential equation (14) given above. The variable element is a thermistor which is heated up to a high temperature, practically instantaneously, by the heater, and then

* This circuit is due to S. Darlington.

¹ Developed by R. F. Wick.

allowed to cool off naturally. By choosing the electrical and thermal constants in the circuit correctly the resulting smoothing can be made to approximate that obtained in a memory point circuit.

As noted earlier, all these variable circuits require some auxiliary control means to reset the variable circuits to zero whenever a new target is engaged or the current target makes a sudden change in course. In the T15 memory point system this function was performed by an operator. The operator was aided by a series of meters which compared the instantaneous memory point rates with average rates set in some time previously by hand. The visual indication of a change in course, calling for the selection of a new memory point, was a relatively large, smoothly and decisively varying deflection on the meters. In contrast, normal tracking errors appeared as relatively small random fluctuations of the needles. The circuits of Figures 7 and 12, which were intended for bombsight applications, were also under the control of an operator, who was supposed to start the mechanism at the beginning of each bombing run.

Two control methods were used for the circuit of Figure 13. In one, large changes in rate, corresponding to probable changes in target

course, were distinguished by comparing the instantaneous value of the target rate, as obtained directly from a differentiator, with the smoothed value obtained at the output of the smoothing circuit. In the other method, equivalent information was obtained by again differentiating the instantaneous value of the target rate, making a second derivative of the target coordinate. In either case this rate difference or second derivative information was used to control a gas tube, which went off, supplying heating current to the variable thermistor, whenever the voltage applied to it exceeded a certain threshold. This threshold evidently marks the minimum change in course for which the variable network will be reset. In order to permit the use of a low threshold, without making the circuit unduly liable to false operation because of the effect of tracking errors, the gas tube input voltage was first transmitted through a low-pass filter which suppressed most of the energy due to tracking errors. A considerable amount of work was done on the proportioning of this filter to provide the best protection against false operation with a low threshold and with minimum delay in resetting in case a change of course actually does occur, but the problem remains an interesting subject for research.

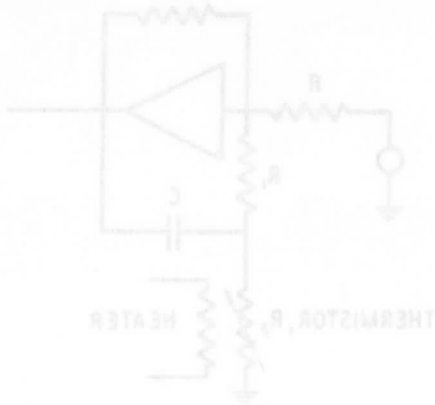


FIGURE 12. Another specific limited range time variable feedback smoothing circuit.



FIGURE 13. Memory point method of obtaining smoothed rate.

corresponds approximately to the second derivative of the differential equation (14) given above. The variable element is a thermistor which is heated up to a high temperature, and then locally instantaneously, by the heater, and then

The memory point method of obtaining smoothed rate is used in the T15 bombsight. In this application, however, it is somewhat more complicated than in the simple illustration described above. This is due to the fact that the position data and the memory point are in the polar coordinate system, whereas the rate components are referred to a tilted and rotating rectangular coordinate system which is determined by the instantaneous line of sight.

* This circuit is due to E. Jamnagor.
† Developed by H. E. Wick.

APPENDIX A

NETWORK THEORY

THIS APPENDIX GIVES a summary of linear network theory which is pertinent to the analysis and design of data-smoothing and prediction circuits. It is incomplete in many respects and should therefore be supplemented by reference to established textbooks on the subject. However, it contains some results which are new.

The present summary will be concerned mainly with fixed linear networks. Variable linear networks will be considered briefly in the last section.

A.1 IMPULSIVE ADMITTANCE

A fixed linear transmission network is one in which the response $V(t)$ is related to the impressed signal $E(t)$ by a linear differential equation of the form

$$b_n \frac{d^n V}{d(t)^n} + b_{n-1} \frac{d^{n-1} V}{d(t)^{n-1}} + \dots + b_0 V = a_m \frac{d^m E}{d(t)^m} + a_{m-1} \frac{d^{m-1} E}{d(t)^{m-1}} + \dots + a_0 E \quad (1)$$

with constant coefficients. It is well-known that the solutions of such a differential equation obey the "superposition principle." This makes it possible to formulate the response of the network to any signal, in terms of its response to certain standard signals.

A convenient standard signal for analytical purposes is the "unit impulse." It may be regarded as the limit of the rectangular pulse shown in Figure 1 as the duration of the pulse

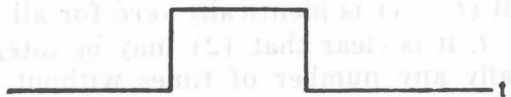


FIGURE 1. Rectangular pulse signal.

is decreased indefinitely while the amplitude is increased in such a way that the area under the pulse is always unity. The limiting function thus defined does not exist in a strict mathematical sense. However, it is very convenient for analytical purposes, and seldom leads to difficulties, to proceed as though the limiting function did exist. An impulse occurring at

$t = \lambda$ is conventionally denoted by the singular function $\delta_0(t - \lambda)$ where

$$\delta_0(\tau) \equiv 0 \quad \text{if } \tau \neq 0$$

$$\int_{-\infty}^t \delta_0(\tau) d\tau \equiv 0 \quad \text{if } t < 0$$

$$\equiv 1 \quad \text{if } t > 0$$

The response of a fixed network to an impulse or any form of signal is independent of the time at which the signal is applied, provided it is expressed as a function of the time relative to the application of the signal. Let $W(t)$ be the response to the signal $\delta_0(t)$. This is called the "impulsive admittance" of the network. Physically, it must be identically zero for negative values of t . For an impulse applied at $t = \lambda$ the response will therefore be $W(t - \lambda)$, which is identically zero for $t < \lambda$.

A physical signal $E(t)$ such as the one shown in Figure 2 may be resolved into an infinite

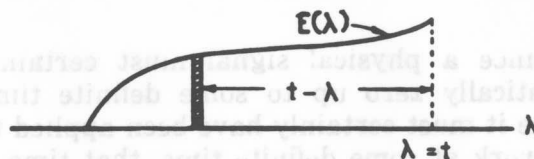


FIGURE 2. Derivation of superposition theorem.

succession of elementary impulses. The strength of the typical elementary impulsive component, such as the one shown in Figure 2 as occurring at time λ , is $E(\lambda)d\lambda$. Its contribution to the response at time t is $E(\lambda) \cdot W(t - \lambda)d\lambda$. Hence the contribution of all the elementary impulsive components of the signal, to the response at time t , is given by the formula

$$V(t) = \int_{-\infty}^{t+} E(\lambda) \cdot W(t - \lambda) d\lambda \quad (2)$$

This is one form of the "superposition theorem" for fixed linear networks.

Before discussing the reasons for the limits of integration indicated in (2), it will be helpful to consider a graphical interpretation other than the one used in deriving the integral. Let $W(t)$ be of the form shown in Figure 3, and let $E(\lambda)$ be of the form shown in Figure 4. To determine the response $V(t)$ at a given value of t , the curve in Figure 3 is turned over from

right to left and placed over the curve in Figure 4 so that its right-hand edge is at $\lambda = t$. The product of the two curves gives a third curve (not shown), which is identically zero for all $\lambda > t$. The area under the third curve is the re-

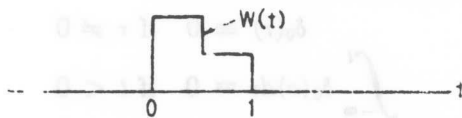


FIGURE 3. An illustrative impulsive admittance.

sponse $V(t)$ at the given value of t . For progressively larger values of t , the curve representing $W(t - \lambda)$ in Figure 4 is simply slid to the right with respect to the curve representing $E(\lambda)$.

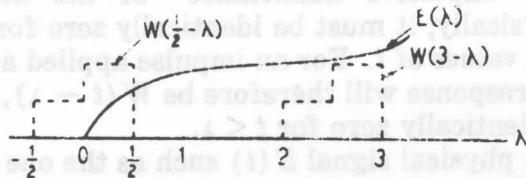


FIGURE 4. Graphical interpretation of superposition theorem.

Since a physical signal must certainly be identically zero up to some definite time, or since it must certainly have been applied to the network at some definite time, that time could be taken arbitrarily as zero and (2) could be written in the form

$$V(t) = \int_0^t E(\lambda) \cdot W(t - \lambda) d\lambda \quad (3)$$

In this form, however, since

$$\int_0^t W(t - \lambda) d\lambda = \int_0^t W(\tau) d\tau$$

is in general a function of t , the response could not be interpreted as a weighted average of the signal. On the other hand, since

$$\int_{-\infty}^t W(t - \lambda) d\lambda = \int_0^{\infty} W(\tau) d\tau$$

is independent of t , the response may be interpreted as a weighted average of the signal, if

$$\int_0^{\infty} W(\tau) d\tau = 1.$$

The necessity of taking the lower limit in (2) as $-\infty$ in order to permit the interpretation of the response as a weighted average of the

signal, is also expressed by the point of view that a fixed network cannot make any physical distinction between having no applied signal and having an applied signal which happens to be of zero amplitude.

Another shortcoming of the form (3) or, for that matter, of the form (2) if we set t as the upper limit of integration, comes from the consideration of impulsive admittances of such a nature that $W(t - \lambda)$ has certain kinds of singularities at $\lambda = t$. For example, the case for direct transmission, expressed in the form

$$V(t) = \int_{-\infty}^t E(\lambda) \cdot \delta_0(t - \lambda) d\lambda$$

is ambiguous because the singularity in the integrand occurs exactly at one end of the range of integration. However, the form

$$V(t) = \int_{-\infty}^{t+} E(\lambda) \cdot \delta_0(t - \lambda) d\lambda$$

leads, without ambiguity, to the result $V(t) = E(t)$. This example is not trivial. Every network which transmits infinite frequency must have an impulsive admittance of such a nature that $W(t - \lambda)$ contains a singularity of the form $\delta_0(t - \lambda)$. Any attempt to rule out such a singularity on the ground that physical networks cannot in fact transmit infinite frequency, complicates the analysis and design of networks unduly. If a network is capable of, or is expected to transmit frequencies at the top of the range of interest or importance, it is simpler to assume that the network is capable of, or is expected to transmit all frequencies above that range.

One other advantage of taking the limits of integration as indicated in (2) may be called to attention. Keeping in mind that $E(\lambda)$ is identically zero for all values of λ below some definite though perhaps unknown value, and that $W(t - \lambda)$ is identically zero for all values of $\lambda > t$, it is clear that (2) may be integrated partially any number of times without incurring the burden of carrying a string of terms outside of the integral. After one partial integration we have

$$V(t) = \int_{-\infty}^t E'(\lambda) \cdot W(t - \lambda) d\lambda \quad (4)$$

where

$$W(t) = \int_0^t W(\tau) d\tau \quad (5)$$

Since $E'(\lambda)$ is identically zero for all values of λ in which $E(\lambda)$ is identically zero, and since

$A(t - \lambda)$ is identically zero for all values of $\lambda > t$, a second partial integration may be performed with no more formal complication than the first partial integration. The fact of the matter is that the terms which ordinarily arise in partial integrations, outside of the integral, are here carried under the integral by singularities of the integrand.

The superposition theorem in the form (4) may be derived directly in a manner similar to the derivation of (2). $A(t - \lambda)$ is the response of the network to a Heaviside unit step function $H(t - \lambda)$ applied at $t = \lambda$, where

$$\begin{aligned} H(t - \lambda) &\equiv 0 & \text{when } t < \lambda \\ &\equiv 1 & \text{when } t > \lambda. \end{aligned}$$

The signal is resolved into an infinite succession of elementary step functions of amplitude $E'(\lambda)d\lambda$ wherever $E(\lambda)$ is continuous, and finite step functions of amplitude $dE(\lambda)$ wherever $E(\lambda)$ has a finite discontinuity. The contribution of each elementary step function to the response at time t is $E'(\lambda) \cdot A(t - \lambda)d\lambda$, that of each finite step function is $A(t - \lambda) \cdot dE(\lambda)$. Hence, the response is given formally by (4) with the understanding that $E'(\lambda)d\lambda$ is to be interpreted as $dE(\lambda)$ wherever $E(\lambda)$ is discontinuous.^a

The response $A(t)$ of the network to a Heaviside unit step function $H(t)$ applied at $t = 0$ is called the "indicial admittance" of the network. It is more familiar, in the field of linear transmission theory, than the impulsive admittance to which it is related by (5), but in this monograph preference is given to the use of the impulsive admittance. In the theory of linear differential equations the impulsive admittance is known as a Green's function.

It is often convenient to express the response so that the variable of integration represents the age of the elementary components of the signal. Introducing the age variable

$$\tau = t - \lambda \quad (6)$$

into (2), we have

$$V(t) = \int_{0-}^{\infty} E(t - \tau) \cdot W(\tau) d\tau. \quad (7)$$

^a Formula (4) may be written in the Stieltjes form

$$V(t) = \int_{-\infty}^{t+} A(t - \lambda) dE(\lambda).$$

Alternatively, we may take the point of view that $E'(\lambda)$ contains impulsive singularities wherever $E(\lambda)$ is discontinuous. This point of view is generalized in Appendix B.

In this form it is clear that the weighting of signal components is on the basis of age only. A fixed network may be said to have a memory which is a function only of the age of past events.

In the preliminary stages of designing a smoothing network, the weighting function $W(\tau)$ is generally prescribed to be identically zero when $\tau > T$ say, as well as when $\tau < 0$. This does not violate the conditions of physical realizability. However, such a weighting function cannot be obtained exactly with a network of a finite number of discrete impedance elements. A finite network invariably yields a weighting function with a "tail" which extends to infinity.

A.2 TRANSMISSION FUNCTION

Theoretically, the impulsive admittance of a prescribed network may be determined directly from the differential equations of the network in a perfectly straightforward manner. Practically, however, it is very difficult to do so if the network has more than two meshes. Furthermore, the technical problem of designing a network directly from a prescribed impulsive admittance is even more difficult, particularly if the impulsive admittance is not exactly realizable.

These difficulties may be avoided by recourse to the highly developed methods of network analysis and synthesis used in the field of communication circuits. These methods are based upon the steady-state properties of networks.

If a signal consisting of the single sinusoid $\cos \omega t$ is applied to an invariable or fixed linear transmission network, the steady-state response^b will also be a single sinusoid of the same frequency. The amplitude and phase of the response, relative to the signal, will in general depend upon the frequency. The response may be regarded as the resultant of an "inphase component" proportional to $\cos \omega t$, and a "quadrature component" proportional to $\sin \omega t$, with amplitude coefficients which are functions of the frequency. Furthermore, since the signal is an even function of the frequency, the response should also be an even function of the frequency.^c Hence, the response will

^b This is the response apart from transient components, assuming that the latter vanish exponentially with time after the signal is impressed.

^c The signal is also an even function of the time but this is due only to the particular choice of origin which is arbitrary.

be of the form $G(\omega^2) \cos \omega t - \omega H(\omega^2) \sin \omega t$, where G and H are even real functions of frequency.

By a suitable shift of the origin of time it follows that if the impressed signal is $\sin \omega t$, the steady-state response will be of the form $G(\omega^2) \sin \omega t + \omega H(\omega^2) \cos \omega t$.

These two results may be combined into a simpler expression without any loss of individuality. Since $e^{i\omega t} = \cos \omega t + i \sin \omega t$ where $i = \sqrt{-1}$, we have

$$V(t) = [G(\omega^2) + i\omega H(\omega^2)] \cdot e^{i\omega t} \quad \text{if } E(t) = e^{i\omega t}.$$

A further simplification may be achieved by replacing $i\omega$ by p , and $G(-p^2) + pH(-p^2)$ by $Y(p)$, so that

$$V(t) = Y(p) \cdot e^{pt} \quad \text{if } E(t) = e^{pt}. \quad (8)$$

$Y(p)$ is called the "steady-state transmission function" or just "transmission function" for short.

Strictly speaking, (8) expresses the relation of steady-state response to signal only if $p = i\omega$. However, it is customarily called a steady-state relation even when p is not a pure imaginary quantity. It may be noted that $Y(p)$ is real when p is real.

The simplicity of steady-state analysis derives from the fact that time occurs in the signal and throughout the network only in the form e^{pt} . In particular, the determination of the transmission function is reduced to the solution of simultaneous algebraic equations which do not involve the time factor. For a network in which the signal and the response are related by the linear differential equation (1) with constant coefficients, we obtain simply

$$Y(p) = \frac{a_0 + a_1 p + \dots + a_m p^m}{b_0 + b_1 p + \dots + b_n p^n}.$$

It may be noted that the poles of the transmission function, also referred to as "infinite-gain points" in the p -plane, correspond to the roots of the characteristic function of the differential equation. Physical restrictions on the location of infinite-gain points will be considered in Section A.9.

A.5 RELATIONSHIP BETWEEN IMPULSIVE ADMITTANCE AND TRANSMISSION FUNCTION

A relationship between the impulsive admittance and the transmission function of a net-

work may be obtained from (7). Putting $E(t) = e^{pt}$ when $t > 0$, we get

$$\begin{aligned} V(t) &= e^{pt} \int_0^t W(\tau) e^{-p\tau} d\tau \\ &= e^{pt} \int_0^\infty W(\tau) e^{-p\tau} d\tau \\ &\quad - e^{pt} \int_t^\infty W(\tau) e^{-p\tau} d\tau. \end{aligned} \quad (9)$$

The second term in (9) is a transient term due to the fact that we have taken $E(t) \equiv 0$ when $t < 0$. The first term in (9), which involves the time only through e^{pt} , is the steady-state term. Comparing this term with (8) we get

$$Y(p) = \int_0^\infty W(t) e^{-pt} dt \quad (10)$$

or, in the notation which will be introduced in the next section

$$Y(p) = L[W(t)]. \quad (11)$$

A.4 LAPLACE AND INVERSE LAPLACE TRANSFORMS

The frequent use which is made of the Laplace transform and its inverse, in the analysis and design of fixed linear networks, warrants a brief discussion of these transforms.

Given a function $f(t)$ which is identically zero when $t < 0$, its Laplace transform $g(p)$ is defined by the formula

$$g(p) = L[f(t)] = \int_0^\infty f(t) e^{-pt} dt. \quad (12)$$

This is usually written with 0 for the lower limit, but by having the point $t = 0$ inside the range of integration, instead of at the end, we secure the same advantages for (12) that we gained in the case of (2) by having the point $\lambda = t$ inside the range of integration. Since $f(t)$ is identically zero when $t < 0$ we could write $-\infty$ for the lower limit in (12), but this would run the risk of confusion with the so-called "bilateral Laplace transform." On the whole, it is worth while to have a constant reminder that functions $f(t)$ which are not identically zero when $t < 0$ are ruled out.

The integral in (12) is usually not convergent for all values of p . That is, in order to secure convergence of the integral, it may be necessary to assume $R(p) > a$, where $R(p)$ is the real part of p , and a is a real number. The

result of the integration is a representation of $g(p)$ in the half-plane $R(p) > a$. Since the representation is analytic throughout the half-plane, the principle of analytic continuation allows us to extend the definition of $g(p)$ to the remainder of the p -plane.

Given a function $g(p)$ which is analytic throughout the half-plane $R(p) \geq c$ where c is a real number, its inverse Laplace transform $f(t)$ is given by the formula

$$f(t) = L^{-1}[g(p)] = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} g(p) e^{pt} dp \quad (13)$$

provided $f(t)$ is identically zero when $t < 0$. If the result of the integration in (13) is not identically zero when $t < 0$, $g(p)$ is not a Laplace transform and the application of the inverse transformation to it is meaningless.

TRANSLATION THEOREM

A useful theorem can be established at this point. This is the *translation theorem*.

If

$$G(p) = L[F(t)]$$

then

$$L^{-1}[G(p)e^{-pa}] = F(t-a)$$

provided that $F(t-a) \equiv 0$ when $t < 0$. Translation is to the right or left according as a is positive or negative.

If it happens that $F(t) \equiv 0$ when $t < t_0$ where $t_0 \geq 0$, then the restriction is that $a \geq -t_0$. That is, a limited amount of translation to the left is permissible. In general, $t_0 = 0$ and the restriction is therefore that $a \geq 0$. This theorem follows readily from (12) or (13).

In all of the applications of (13) which we have any occasion to make in the analysis and design of fixed linear networks, the function $g(p)$ may be resolved into a sum of terms of the form $G(p)e^{-pa}$ where $a \geq 0$ and $G(p)$ is a rational algebraic function with real coefficients. Making use of the translation theorem, the problem of evaluating $L^{-1}[g(p)]$ reduces to that of evaluating $L^{-1}[G(p)]$. Now, $G(p)$ may be resolved into a sum of terms of the form p^m or $1/(p-a)^{m+1}$ where $m = 0, 1, 2, \dots$. We shall consider these two cases separately.

The case $G(p) = p^m$ will be treated by means of (12) and some limiting processes. In Section A.1 the unit impulse was regarded as the limit of a rectangular pulse of duration T and amplitude $1/T$. By means of (12) the Laplace

transform of such a pulse over the interval $0 \leq t \leq T$ is

$$\frac{1 - e^{-pT}}{pT}$$

Hence

$$L[\delta_0(t)] = \lim_{T \rightarrow 0} \frac{1 - e^{-pT}}{pT} = 1.$$

Formally therefore

$$L^{-1}[1] = \delta_0(t) \quad (14)$$

Similarly, the Laplace transform of a pulse over the interval $a \leq t \leq a+T$ where $a > 0$ is

$$\frac{1 - e^{-pT}}{pT} e^{-pa}$$

Hence

$$L[\delta_0(t-a)] = \lim_{T \rightarrow 0} \frac{1 - e^{-pT}}{pT} e^{-pa} = e^{-pa}.$$

Formally therefore

$$L^{-1}[e^{-pa}] = \delta_0(t-a).$$

The last result follows directly from (14) using the translation theorem.

Next, let

$$\delta_1(t) = \lim_{T \rightarrow 0} \frac{\delta_0(t) - \delta_0(t-T)}{T}.$$

This is the limiting case, as shown in Figure 5, of two impulses of strengths $1/T$ and $-1/T$ separated by a time interval T . It may be called

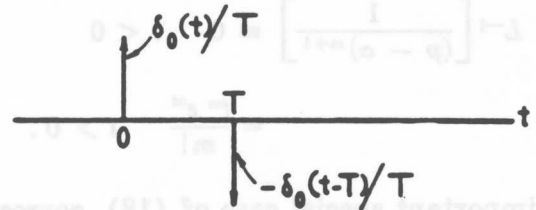


FIGURE 5. An impulse doublet.

an impulse of second order. By (12) and the previous results

$$L[\delta_1(t)] = \lim_{T \rightarrow 0} \frac{1 - e^{-pT}}{T} = p.$$

Formally therefore

$$L^{-1}[p] = \delta_1(t). \quad (15)$$

Proceeding in this fashion we may define an impulse of $(m+1)$ th order as

$$\delta_m(t) = \lim_{T \rightarrow 0} \frac{\delta_{m-1}(t) - \delta_{m-1}(t-T)}{T} \quad (16)$$

and we may then show that

$$L[\delta_m(t)] = p^m.$$

Formally therefore

$$L^{-1}[p^m] = \delta_m(t). \quad (17)$$

This disposes of the case $G(p) = p^m$ where $m = 0, 1, 2, \dots$

The case $G(p) = 1/(p - \alpha)^{m+1}$ will be treated by means of (13) and Jordan's lemma.

JORDAN'S LEMMA

If all the singularities of $G(p)$ can be enclosed by a circle of finite radius with center at the origin, and if $G(p) \rightarrow 0$ uniformly with respect to $\arg z$ as $|z| \rightarrow \infty$, then

$$\lim_{r \rightarrow \infty} \left[\int_{\Gamma} G(p) e^{pt} dp \right] = 0$$

where Γ is a semicircle of radius ρ , with center at the origin, to the right of the imaginary axis if t is negative, to the left of the imaginary axis if t is positive.

By the use of this lemma the contour of integration in (13) may be closed and the integration may then be performed by the method of residues. In the case

$$G(p) = \frac{1}{(p - \alpha)^{m+1}} \quad \text{where } m = 0, 1, 2, \dots$$

we readily obtain

$$L^{-1} \left[\frac{1}{(p - \alpha)^{m+1}} \right] \equiv 0 \quad t < 0$$

$$= \frac{t^m e^{\alpha t}}{m!} \quad t > 0. \quad (18)$$

An important special case of (18), corresponding to $\alpha = 0$, is

$$L^{-1} \left[\frac{1}{p^{m+1}} \right] = \frac{t^m}{m!} \quad t > 0. \quad (19)$$

Another useful theorem which is readily established by means of (12) and (13) is Borel's theorem.

BOREL'S THEOREM

If $g(p)$, $g_1(p)$, $g_2(p)$ are the Laplace transforms of $f(t)$, $f_1(t)$, $f_2(t)$, respectively, and if

$$g(p) = g_1(p) g_2(p)$$

then

$$f(t) = \int_{0-}^{t+} f_1(t - \lambda) \cdot f_2(\lambda) d\lambda$$

$$= \int_{0-}^{t+} f_1(\tau) \cdot f_2(t - \tau) d\tau.$$

The functions $f_1(t)$ and $f_2(t)$ are subject to conditions which permit the inversion of the order of integration in the following proof. However, these conditions are seldom of any concern. We have

$$f(t) = L^{-1}\{g_1(p) \cdot L[f_2(t)]\}$$

$$= \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \left[g_1(p) e^{pt} \int_{0-}^{\infty} f_2(\lambda) e^{-p\lambda} d\lambda \right] dp.$$

Inverting the order of integration and noting that

$$\frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} g_1(p) e^{p(t-\lambda)} dp = 0 \quad \text{if } \lambda > t$$

$$= f(t - \lambda) \quad \text{if } \lambda < t$$

we obtain the result stated in the theorem.

A.5 ALTERNATIVE EXPRESSION OF THE RESPONSE-TO-SIGNAL RELATIONSHIP

The result (8) obtained in Section A.2 suggests an operational expression of the form

$$V(t) = Y(p) \cdot E(t) \quad (20)$$

for the response-to-signal relationship whatever the signal $E(t)$ might be. If the equivalence of this operational expression to (2) is taken as a matter of definition we may readily discover the nature of the implied operation.

In the light of Borel's theorem, (2) may be expressed in the form

$$L[V(t)] = L[W(t)] \cdot L[E(t)]$$

under the permissible assumption that $E(t) = 0$ when $t < 0$. Hence

$$V(t) = L^{-1}\{L[W(t)] \cdot L[E(t)]\}$$

or, by (11)

$$V(t) = L^{-1}\{Y(p) \cdot L[E(t)]\}. \quad (21)$$

This is, therefore, in general the meaning of the operational expression (20).⁴

⁴ We note that if $S(p) = L[E(t)]$, the operational expression

$$V(t) = S(p) \cdot W(t)$$

is equivalent to (20). This form is used in Section 10.4 and in Appendix B.

The symmetry of the impulsive admittance is expressed by

$$W(T - t) = W(t)$$

Since $W(t) \equiv 0$ when $t < 0$, it must be so also when $t > T$. Hence

$$Y(p) = \int_{0-}^{T/2} W(t)e^{-pt} dt + \int_{T/2}^{T+} W(t)e^{-pt} dt.$$

By a change of variable of integration the second term may be expressed in the form

$$\int_{0-}^{T/2} W(T - t)e^{-p(T-t)} dt$$

or, because of the symmetry, in the form

$$e^{-pT} \int_{0-}^{T/2} W(t)e^{pt} dt.$$

Hence, if the first term in $Y(p)$ be denoted by

$$Y_1(p) = \int_{0-}^{T/2} W(t)e^{-pt} dt$$

we have

$$\begin{aligned} Y(p) &= Y_1(p) + Y_1(-p)e^{-pT} \\ &= [Y_1(p)e^{pT/2} + Y_1(-p)e^{-pT/2}]e^{-pT/2}. \end{aligned}$$

At real frequencies ($p = i\omega$) the bracketed factor is evidently an even real function of ω . Hence

$$Y(i\omega) = Q(\omega^2) \cdot e^{-i\omega T/2}. \quad (24)$$

Apart from discontinuities in the phase angle of the transmission function at real frequencies ω for which $Q(\omega^2)$ is zero, the phase angle is proportional to frequency. Such a transmission function is referred to as a linear phase transmission function. Sinusoidal components of the signal, of frequencies less than the lowest frequency at which $Q(\omega^2)$ vanishes, suffer phase retardations in transmission in proportion to their frequencies. These components therefore contribute no delay distortion. They are delayed by a uniform amount, just as they are in a properly terminated distortionless, uniform transmission line, although in the case of (24) they contribute amplitude or loss distortion through $Q(\omega^2)$. The delay in (24) is just half of the "smoothing time" T .

4.3 SERIES RELATIONSHIPS BETWEEN IMPULSIVE ADMITTANCE AND TRANSMISSION FUNCTION

Two useful series relationships between impulsive admittances and transmission functions will be derived in this section.

Assume that $W(t)$ admits the series expansion

$$W(t) = A_0 + A_1 t + \dots + \frac{A_m t^m}{m!} + \dots \quad (25)$$

for small positive values of t . Then by (11) and (19)

$$Y(p) = \frac{A_0}{p} + \frac{A_1}{p^2} + \dots + \frac{A_m}{p^{m+1}} + \dots \quad (26)$$

If $A_0 \neq 0$ the transmission cannot drop off faster than 6 db per octave as the frequency increases indefinitely. If the transmission is to drop off ultimately at the rate of $6k$ db per octave all of the A 's up to and including A_{k-2} must be zero. This is to say that the impulsive admittance and all of its derivatives of orders up to and including the $(k-2)$ th must vanish at $t = 0$.

Next, let us suppose that the impulsive admittance and all of its derivatives of orders up to and including the $(k-2)$ th are continuous through all values of t including $t = 0$ except that the $(k-2)$ th derivative is discontinuous only at $t = a$. We may resolve the impulsive admittance into the sum $W_1(t) + W_2(t)$ where $W_1(t)$ and all of its derivatives of orders up to and including the $(k-2)$ th are continuous through all values of t including $t = 0$, while $W_2(t) \equiv 0$ for all values of $t < a$. Then, for small positive values of $t - a$

$$W_2(t) = \frac{A_{k-2}(t-a)^{k-2}}{(k-2)!} + \dots \quad (A_{k-2} \neq 0)$$

whence

$$Y_2(p) = \left(\frac{A_{k-2}}{p^{k-1}} + \dots \right) e^{-pa}.$$

Hence the transmission cannot drop off ultimately faster than $6(k-1)$ db per octave. We may summarize these results in the asymptotic loss theorem.

ASYMPTOTIC LOSS THEOREM.

If the transmission is to drop off ultimately at the rate of $6k$ db per octave as the frequency increases indefinitely, the impulsive admittance and all of its derivatives of orders up to and including the $(k-2)$ th must be continuous through all values of t including $t = 0$.

Discontinuities in $W(t)$ or in some derivative of $W(t)$ cannot occur except at $t = 0$ in the case of physical lumped element networks. Practically, however, rapid changes in $W(t)$

or in some derivative of $W(t)$, at any value of t , may be expected to be associated with much the same behavior of the transmission at reasonably high frequencies. As an example consider the case

$$W(t) = e^{-\alpha t} - e^{-\beta t} \quad (\beta > \alpha > 0).$$

$$Y(p) = \frac{\beta - \alpha}{(p + \alpha)(p + \beta)}.$$

$W(t)$ is continuous through $t = 0$ as long as β is finite but becomes discontinuous there in the limit as $\beta \rightarrow \infty$. The first derivative of $W(t)$ is discontinuous through $t = 0$ even when β is finite. The ultimate slope of the transmission is 12 db per octave, in accordance with the asymptotic loss theorem, but in the range $\alpha < \omega < \beta$ the transmission appears to have a slope of only 6 db per octave.

The importance of the observations made in the preceding paragraph, in the design of a network, is that if we attempt to approximate a $W(t)$ which has a discontinuity in a derivative of lower order at $t = a$ than at $t = 0$, the fact that the physical approximation must have continuous derivatives of all orders and through all values of t except $t = 0$ is not very significant. The ultimate slope of the transmission may not be reached until the frequency is too high to be of any importance.

Another useful relationship between impulsive admittance and transmission function fol-

lows from the assumption that $\int_{0-}^{\infty} t^m W(t) dt$

is finite for $m = 0, 1, 2, \dots$. If we expand the exponential in

$$Y(p) = \int_{0-}^{\infty} W(t)e^{-pt} dt$$

into a power series in pt we get

$$Y(p) = M_0 - M_1 p + \frac{M_2 p^2}{2!} - \frac{M_3 p^3}{3!} + \dots \quad (27)$$

where

$$M_m = \int_{0-}^{\infty} t^m W(t) dt. \quad (28)$$

The quantity M_m is the m th moment of the impulsive admittance.

When $M_0 = 1$ we speak of the response of the network as a weighted average of the impressed signal, and speak of the impulsive admittance $W(t)$ as the weighting function.

A.9 PHYSICAL RESTRICTIONS ON THE TRANSMISSION FUNCTION

The transmission function $Y(p)$ of a lumped element network is a rational algebraic function of p . It is real for real values of p (A.2). Hence, the coefficients must be real, and therefore the roots and poles must either be real or occur in conjugate complex pairs.

Such a function may be expanded into the sum of a polynomial and a rational function whose numerator is of lower degree than the denominator. The latter may therefore be properly expanded into partial fractions. For a partial fraction of the form

$$\frac{1}{(p - \alpha)^m} \quad \text{where } m = 1, 2, \dots$$

the contribution to the impulsive admittance $W(t)$ is by (18)

$$L^{-1} \left[\frac{1}{(p - \alpha)^m} \right] = \frac{t^{m-1}}{(m-1)!} e^{\alpha t} \quad (t > 0).$$

For a pair of partial fractions of the form

$$\frac{A + iB}{(p - \alpha + i\beta)^m} + \frac{A - iB}{(p - \alpha - i\beta)^m}$$

the contribution to the impulsive admittance is

$$\frac{2t^{m-1}}{(m-1)!} e^{\alpha t} (A \cos \beta t + B \sin \beta t).$$

Since the impulsive admittance is the response to an impulsive signal it is clear that for a stable network the impulsive admittance must be free of terms which increase indefinitely with time, either on account of an amplitude factor of the form $e^{\alpha t}$ where $\alpha > 0$, or, in the event that $\alpha = 0$, on account of an amplitude factor of the form t^{m-1} where $m > 1$. Hence, the physical restrictions on the transmission function are:

1. No poles with positive real parts.
2. Poles on the imaginary p axis must be simple.*

The poles of a passive transmission function correspond to modes of free motion.^{15b} Each of them may be shown^{15a} to satisfy an equation of the form

$$pT + F + \frac{V}{p} = 0$$

where T, F, V are positive quantities whose values depend upon the particular mode and

* Poles on the imaginary p axis must also be ruled out on the ground that persistent transients cannot be tolerated any more than growing transients.

its activity. However, T is zero in the absence of kinetic energy, F is zero in the absence of energy dissipation, and V is zero in the absence of potential energy. It follows that in the absence of coils or in the absence of condensers, the transmission function must have poles only on the negative real p axis.

For extremely narrow-band, low-pass applications, such as data smoothing, it is not practicable to build networks which call for coils because these generally turn out to be of many thousands of henries in inductance. The exclusion of coils from these applications does not, however, rule out transmission functions with complex poles. These may be realized with RC networks in feedback amplifier circuits as is shown in Chapter 12.

A.10

QUASI-DISTORTIONLESS TRANSMISSION NETWORKS

A quasi-distortionless transmission network is one which is distortionless only in a certain sense. This sense will be made clear in this section.

Let

$$Y(p) = \frac{1 + a_1 p + a_2 p^2 + \dots + a_n p^n}{1 + b_1 p + b_2 p^2 + \dots + b_n p^n} \quad (29)$$

This may also be written in the form

$$Y(p) = 1 + c_1 p + \frac{c_2 p^2}{2!} + \dots + \frac{c_r p^r}{r!} + p^{r+1} g(p). \quad (30)$$

Obviously $g(p)$ will be a rational function with the same denominator as $Y(p)$ and a numerator of $(n-1)$ th degree. If we now apply a signal of the form

$$E(t) = 0 \quad \text{for } t < 0 \\ = t^r \quad \text{for } t > 0$$

the response, by (21), will be

$$V(t) = t^r + r c_1 t^{r-1} + \frac{r!}{2!(r-2)!} c_2 t^{r-2} + \dots + c_r + r! L^{-1}[g(p)] \quad (t > 0).$$

If the coefficients in the rational expression for $Y(p)$ are such that

$$c_1 = t_f, c_2 = t_f^2, \dots, c_r = t_f^r \quad (31)$$

then

$$V(t) = (t + t_f)^r + r! L^{-1}[g(p)] \quad (t > 0). \quad (32)$$

The second term vanishes exponentially with time. The first term is an advanced or a retarded facsimile of the applied signal accord-

ing to whether t_f is positive or negative. We shall say that $Y(p)$ is the transmission function of a network which is quasi-distortionless to the signal t^r .

Obviously a transmission network which is quasi-distortionless to the signal t^r must also be quasi-distortionless to every signal t^s where s is a positive integer less than r , including zero. Hence we may state the quasi-distortionless transmission theorem.

QUASI-DISTORTIONLESS TRANSMISSION THEOREM

If the signal

$$E(t) = 0 \text{ for } t < 0 \\ = \text{polynomial of degree } r \text{ at most in } t \text{ for } t > 0$$

is applied to a "quasi-distortionless transmission network of order r ," the response will be of the form

$$V(t) = E(t + t_f) + 0(e^{-t}) \quad \text{for } t > 0,$$

where $0(e^{-t})$ stands for terms which vanish exponentially with time.

If $t_f > 0$ the transmission network is a predictor for polynomials of degree r at most. However, it does not begin to predict properly until some time has elapsed after the start of the signal, or of a new analytic segment of the signal; that is, until the transients have subsided sufficiently.

If $t_f = 0$ the transmission network may be regarded as a delay-corrected smoother for polynomials of degree r at most. This is obtained simply by taking

$$a_1 = b_1, a_2 = b_2, \dots, a_r = b_r \quad (33)$$

in (29).

A.11 VARIABLE LINEAR NETWORKS

A variable linear transmission network is one in which the response $V(t)$ is related to the impressed signal $E(t)$ by the linear differential equation (1) with coefficients which are prescribed functions of t . The solutions of such a differential equation also obey the superposition principle. Thus it is possible in this case also to formulate the response of the network to any signal in terms of its response to a standard impulsive signal.

The response of a variable network to an impulse or any form of signal depends, how-

ever, on the time at which the signal is applied. For an impulsive signal applied at time λ the response at time t will be represented by $W(t, \lambda)$. This is still called the "impulsive admittance." In the theory of linear differential equations it is known as a Green's function. Physically, it must be identically zero for $t < \lambda$.

The superposition theorem may now be written in the form

$$V(t) = \int_{0-}^{t+} E(\lambda) \cdot W(t, \lambda) d\lambda \quad (34)$$

provided the network has been properly designed and set into operation at $t = 0$. If

$$\int_{0-}^{t+} W(t, \lambda) d\lambda = 1$$

for all values of $t > 0$, the response may be interpreted as a weighted average of the signal. We note that in order to interpret the response as a weighted average of the signal, it is now no longer necessary to take the lower limit in (34) as $-\infty$, as it was in the case of (2) for a fixed network. In other words, a variable network can be designed and set into operation at any time so that components of the signal which arrive before that time are completely ignored.

The analysis and design of variable linear networks are in general much more difficult

than those of fixed linear networks. This is due largely to the fact that there does not yet exist a technique corresponding to the steady-state and operational methods used in connection with fixed networks. However, there is a class of variable networks whose analysis and design are greatly facilitated by the fact that they are related to fixed networks by a transformation of the time variable.

Consider the linear differential equation

$$b_n \frac{d^n V}{dz^n} + b_{n-1} \frac{d^{n-1} V}{dz^{n-1}} + \dots + b_1 \frac{dV}{dz} + V = E$$

with constant coefficients. With appropriate restrictions on the roots of the characteristic function

$$b_n \lambda^n + b_{n-1} \lambda^{n-1} + \dots + b_1 \lambda + 1$$

it represents the response-to-signal relationship in a fixed network, if z is proportional directly to time. However, if z is a more general function of the time, it will correspond to a variable network. The kind of transformation which is desired here is one which transforms the range $-\infty < z < +\infty$ into the range $0 < t < +\infty$ with a one-to-one correspondence. Thus, we may take $z = \log \theta(t)$ where $\theta(t)$ is a positive monotonic increasing function of t in the range $0 < t < +\infty$, with $\lim_{t \rightarrow 0} \theta(t) = 0$. Several examples of $\theta(t)$, including $\theta(t) \equiv t$, are considered in detail in Chapter 14.

THEORETICAL MODIFICATIONS OF SMOOTHING FUNCTIONS TO FIT
NONUNIFORM NOISE SPECTRA

BEST SMOOTHING or weighting functions have been determined in Chapters 10 and 11 under the assumption of random noise with flat spectrum. It has not been worth while in practice to base the choice of best weighting functions on any more elaborate considerations of actual noise spectra, for at least three reasons:

1. The effectiveness of a smoothing network shape of the weighting function.

2. Noise spectra are subject to variations, due to factors which it is not desirable in practice to attempt to control.

3. Elaborate smoothing functions require elaborate networks with close tolerances on element values.

Nevertheless, the theory of smoothing presented in this monograph would not be complete without showing how more general shapes of noise spectra can be considered. Two methods are presented here, which are generalizations of those presented in Sections 10.3 and 10.4, respectively.

A.1 PHILLIPS AND WEISS THEORY*

Let $g(t)$ be the tracking error, and $W(t)$ the impulsive admittance of a smoothing and prediction circuit with smoothing time T . Then the error in prediction due to tracking error only, is

$$V(t) = \int_0^T g(t - \tau) \cdot W(\tau) d\tau.$$

The impulsive admittance $W(\tau)$ will depend also upon the time of flight which, for purposes of analysis, is assumed to be constant. The mean square error is then

$$\begin{aligned} V^2 &= \lim_{L \rightarrow \infty} \frac{1}{2L} \int_{-L}^L V^2(t) dt \\ &= \int_0^T \int_0^T W(\tau_1) \cdot C(\tau_1 - \tau_2) \cdot W(\tau_2) d\tau_1 d\tau_2 \end{aligned}$$

where

$$C(x) = \lim_{L \rightarrow \infty} \frac{1}{2L} \int_{-L}^L g(\lambda) \cdot g(\lambda + x) d\lambda \quad (1)$$

$C(x)$ is the autocorrelation of the error time-function $g(\lambda)$.

For an n th order smoothing and prediction circuit \bar{V}^2 is now minimized with respect to the impulsive admittance under the restrictions*

$$\int_0^T \tau^m W(\tau) d\tau = (-t_f)^m \quad (m = 0, 1, 2 \dots n). \quad (2)$$

Hence $W(\tau)$ must satisfy the integral equation

$$\int_0^T C(t - \tau) \cdot W(\tau) d\tau = k_0 + k_1 t + \dots + k_n t^n \quad (0 \leq t \leq T)$$

where the k_m are constants to be determined. Now, if

$$\int_0^T C(t - \tau) \cdot W_m(\tau) d\tau = t^m \quad (0 \leq t \leq T) \quad (m = 0, 1, 2 \dots n) \quad (3)$$

then

$$W(\tau) = k_0 W_0(\tau) + k_1 W_1(\tau) + \dots + k_n W_n(\tau). \quad (4)$$

The procedure is then to determine $C(x)$ from (1), the $W_m(\tau)$ from (3), the k_m from (2) and (4), and finally $W(\tau)$ from (4). It may be noted that, in general, every k_m will be a polynomial of n th degree in t_f . Hence the $W_m(\tau)$ appearing here are not the same as those defined in Chapter 11, although $W(\tau)$ should be the same if the same $W_0(\tau)$ is used in Chapter 11.

A difficulty of the theory given above is in the solution of the integral equations (3). This difficulty is avoided in the theory given in the next section. However, the integral equations are easily solved in case of flat random noise, when $C(x)$ is simply an impulse of strength K say, at $x = 0$. Then

$$W_m(\tau) = \frac{\tau^m}{K} \quad 0 < \tau < T.$$

Since the strength is irrelevant, it may be taken equal to T so that $W_0(\tau)$ will be normalized.

* These follow from the discussions in Sections A.8 and A.10, especially equations (27), (28), (30), and (31).

For a linear prediction circuit it is then found that

$$W(\tau) = 2 \left(2 + \frac{3t_f}{T} \right) W_0(\tau) - \frac{6}{T} \left(1 + \frac{2t_f}{T} \right) W_1(\tau).$$

Putting $T = 1$ this may be expressed as

$$W(\tau) = w_0(\tau) + G_1(-t_f) w_1^{(1)}(\tau)$$

in terms of the $G_m(\tau)$ and $W_m(\tau)$ of Section 11.3.

2.3 SYMMETRY OF BEST SMOOTHING FUNCTIONS

The theory of Phillips and Weiss offers the most direct proof that the best smoothing or weighting function must be symmetrical, regardless of the noise power spectrum. The situation is that of minimizing (1) under only one of the restrictions (2), viz., the normalizing condition

$$\int_{-T}^0 W(\tau) d\tau = 1 \quad (5)$$

The weighting function is therefore determined, up to a constant scale factor, by the condition that

$$\int_0^T C(t - \tau) \cdot W(\tau) d\tau = k, \quad (6)$$

where k is a constant. Substituting $T - t$ for t and $T - \tau$ for τ , we have

$$\int_0^T C(\tau - t) \cdot W(T - \tau) d\tau = k. \quad (7)$$

Since $C(-x) = C(x)$, and since $W(\tau)$ is determined uniquely by (6) and (5), it follows from (6) and (7) that

$$W(T - \tau) = W(\tau). \quad (8)$$

2.3 GENERALIZATION OF ELEMENTARY PULSE METHOD

The noise power transmitted through a network may be expressed in the familiar form

$$P = \int_0^\infty N(\omega^2) \cdot |Y(i\omega)|^2 d\omega$$

where $N(\omega^2)$ is the noise power spectrum and $Y(p)$ is the transmission function of the network. Assuming that $N(\omega^2)$ is a rational function of ω^2 , which is finite at all finite values of ω including zero, it is possible to determine a

rational function $S(p)$, which has no poles on or to the right of the imaginary axis in the p -plane with the exception of the point at infinity, and such that

$$|S(i\omega)|^2 = N(\omega^2).$$

It may be readily shown that

$$P = \pi \int_{-\infty}^{\infty} [F(t)]^2 dt \quad (9)$$

where $F(t)$ is related to the impulsive admittance $W(t)$ by the operational equation

$$F(t) = S(p) \cdot W(t) \quad (10)$$

The problem is now to minimize (9) under the restriction

$$\int_{-\infty}^{\infty} W(t) dt = 1 \text{ when } t_0 > 1. \quad (11)$$

Let

$$S(p) = k \frac{Q(p)}{R(p)}$$

where

$$\frac{Q(p)}{R(p)} = \frac{(p + \alpha_1)(p + \alpha_2) \cdots (p + \alpha_m)}{(p + \beta_1)(p + \beta_2) \cdots (p + \beta_n)}$$

and k is of no consequence. One or more of the α 's, but none of the β 's may be zero. Since the existence of the integral in (9) imposes the requirement that $F(t)$ have no discontinuities of higher type than finite jumps in the range $0 < t < \infty$, the continuity conditions on $W(t)$ in (10) must depend upon the difference between m and n in the expressions for $Q(p)$ and $R(p)$.

If $m \geq n$, it is fairly obvious that $W(t)$ must be differentiable, in the ordinary sense, exactly $m - n$ times. In other words, $W(t)$ and all its derivatives up to and including the $(m - n - 1)$ th must be continuous, but the $(m - n)$ th derivative may have finite jumps. If $m < n$ we must consider the introduction into $W(t)$ of discontinuities of higher type than finite jumps. These discontinuities arise in the formal extension of the concept of differentiation to functions containing finite jumps.

If a function $\phi(t)$ has a finite jump of amplitude A_0 at $t = a$, the value of $\phi'(t)$ at that point will be indicated formally as $A_0 \cdot \delta_0(t - a)$ where $\delta_0(t - a)$ is a unit impulse at $t = a$. If $\phi'(a + 0) - \phi'(a - 0) = A_1$, the value of $\phi''(t)$ at $t = a$ will be indicated formally as $A_0 \cdot \delta_1(t - a) + A_1 \cdot \delta_0(t - a)$ where $\delta_1(t - a)$ is a

unit doublet at $t = a$. And so on, for higher derivatives of $\phi(t)$.

The expression (9) is a minimum under the restriction (11) if $W(t)$ satisfies the differential equation

$$Q(p) \cdot Q(-p) \cdot W(t) = \text{const.} \quad (12)$$

when $0 < t < 1$ and $Y(p)$ the condition

$$\frac{1}{2\pi i} \int_{-i\infty}^{i\infty} S(p) \cdot S(-p) \cdot Y(p) e^{pt} dp = \text{const.} \quad (13)$$

when $0 < t < 1$.

The restriction (11) itself requires that $W(t) = 0$ when $t > 1$, and

$$\int_0^{1+} W(t) dt = 1. \quad (14)$$

CASE I. ($n = 0$)

The general solution of (12) contains $2m + 1$ constants of integration which are determined by (14) and the $2m$ continuity conditions that $W(t)$ and all of its derivatives up to and including the $(m - 1)$ th must vanish at $t = 0$ and $t = 1$.

CASE II. ($n \neq 0, m \geq n$)

The general solution of (12) contains $2m + 1$ constants of integration which are reduced to $2n$ in number by (14) and the $2(m - n)$ continuity conditions that $W(t)$ and all of its derivatives up to and including the $(m - n - 1)$ th must vanish at $t = 0$ and at $t = 1$. The remaining $2n$ constants are determined by (13).

The left-hand member of (13) may be formulated by the method of residues. The expression for $Y(p)$ should first be separated into two parts so that

$$Y(p) = Y_L(p) + Y_R(p)e^{-p}$$

where $Y_L(p)$ and $Y_R(p)$ are rational functions of $S(p) \cdot S(-p) \cdot Y_L(p) e^{pt}$ in the left-hand half of the p -plane for the first part of $Y(p)$, and in the right-hand half for the second part. Hence, if the sum of the residues of $S(p) \cdot S(-p) \cdot Y_L(p) e^{pt}$ in the left-hand half of the p -plane be denoted by Σ_L , and if the sum of the residues of $S(p) \cdot S(-p) \cdot Y_R(p) \cdot e^{p(t-1)}$ in the right-hand half of the p -plane be denoted by Σ_R , then the condition (13) reduces to

$$\Sigma_L - \Sigma_R = \text{const.} \quad (15)$$

CASE III. ($n \neq 0, m < n$)

The $2m + 1$ constants of integration in the general solution of (12) are first increased to $2n + 1$ by appending the $2(n - m)$ singularities

$$\delta_0(t), \quad \delta_1(t), \quad \dots \delta_{n-m-1}(t) \\ \delta_0(t-1), \delta_1(t-1), \dots \delta_{n-m-1}(t-1)$$

and then reduced to $2n$ by (14). The remainder are determined by (13) or (15).

In formulating

$$Y(p) = L[W(t)]$$

it may be noted that

$$L[\delta_n(t-a)] = p^n e^{-ap} \quad (a \geq 0).$$

EXAMPLE OF CASE I

Let $S(p) = p^m$. The differential equation (12) requires $W(t)$ to be a polynomial of degree $2m$. The conditions at $t = 0$ require it to have a factor t^m , and those at $t = 1$, a factor $(1 - t)^m$. This leaves only (14) to be satisfied. Hence

$$W(t) = \frac{(2m+1)!}{(m!)^2} [t(1-t)]^m \quad (0 \leq t \leq 1)$$

in agreement with (8) of Section 10.8.

EXAMPLE OF CASE II

$$\text{Let} \quad S(p) = \frac{p + \alpha}{p + \beta}. \quad (12)$$

Then, by

$$W(t) = A_0 + A_1 e^{-at} + A_2 e^{at} \quad (0 \leq t \leq 1).$$

Hence

$$Y(p) = \frac{A_0}{p} + \frac{A_1}{p + \alpha} + \frac{A_2}{p - \alpha} \\ - \left[\frac{A_0}{p} + \frac{A_1 e^{-a}}{p + \alpha} + \frac{A_2 e^a}{p - \alpha} \right] e^{-p}$$

$$\Sigma_L =$$

$$\frac{A_0 \alpha^2}{\beta^2} - \left[A_0 \frac{\alpha^2 - \beta^2}{2\beta^2} - A_1 \frac{\alpha + \beta}{2\beta} + A_2 \frac{\alpha - \beta}{2\beta} \right] e^{-a}$$

$$\Sigma_R = \left[A_0 \frac{\alpha^2 - \beta^2}{2\beta^2} + A_1 \frac{\alpha - \beta}{2\beta} e^{-a} - A_2 \frac{\alpha + \beta}{2\beta} e^a \right] e^{a(t-1)}.$$

Condition (15) is satisfied if

$$A_1 = \frac{1}{2} A_0 Q e^{a/2} \quad A_2 = \frac{1}{2} A_0 Q e^{-a/2}$$

where

$$Q = \frac{\alpha_2 - \beta_2}{\beta \left(\alpha \sinh \frac{\alpha}{2} + \beta \cosh \frac{\alpha}{2} \right) \alpha^2}.$$

Hence

$$W(t) = \frac{1 + Q \cosh \alpha \left(t - \frac{1}{2} \right)}{1 + \frac{2Q}{\alpha} \sinh \frac{\alpha}{2}} \quad (0 \leq t \leq 1).$$

In the limit as $\alpha \rightarrow 0$, $S(p) = \frac{p}{p + \beta}$

and

$$W(t) = \frac{1 + \frac{\beta^2}{2 + \beta} t(1 - t)}{1 + \frac{1}{6} \frac{\beta^2}{2 + \beta}} \quad (0 \leq t \leq 1).$$

In terms of expressions (12), Section 11.3.

$$W(t) = \frac{W_0(t) + k w_1(t)}{1 + k} \quad (0 \leq t \leq 1)$$

where $k = 1/6[\beta^2/(2 + \beta)]$. This is reminiscent of Stibitz's results mentioned in Section 10.3.

EXAMPLE OF CASE III

Let $S(p) = 1/1 + \beta$. Then, by (12) and the rule for appending singularities in Case III

$$W(t) = A_0 + A_1 \delta_0(t) + A_2 \delta_0(t - 1) \quad (0 \leq t \leq 1).$$

Hence

$$Y(p) = \frac{A_0 + A_1 p}{p} - \frac{A_0 - A_2 p}{p} e^{-p}$$

$$\Sigma_L = -\frac{A_0}{\beta^2} + \frac{A_0 - \beta A_1}{2\beta^2} e^{-\mu}$$

$$\Sigma_R = -\frac{A_0 - \beta A_2}{2\beta^2} e^{\mu(1-1)}.$$

Condition (15) is satisfied if

$$A_1 = A_2 = \frac{A_0}{\beta}.$$

Hence

$$W(t) = \frac{1 + \frac{\delta_0(t) + \delta_0(t - 1)}{\beta}}{1 + \frac{2}{\beta}} \quad (0 \leq t \leq 1)$$

PART II

1. *The Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications*, Norbert Wiener, OSRD 870, Report to the Services 19, Research Project DIC-6087, The Massachusetts Institute of Technology, Feb. 1, 1942.
Div. 7-313.1-M2
- 1a. *Ibid.*, Chapter 1.
2. *The Analysis and Design of Servomechanisms*, Herbert Harris, Jr., OSRD 454, Progress Report to the Services 23, The Massachusetts Institute of Technology.
Div. 7-321.1-M7
3. *Behavior and Design of Servomechanisms*, Gordon S. Brown, OSRD 39, Progress Report 2, The Massachusetts Institute of Technology, November 1940.
Div. 7-321.1-M1
4. *Antiaircraft Director T-15*, OEMsr-353, Report to the Services 62, Western Electric Company, Inc., August 1943.
Div. 7-112.2-M5
5. *The Analysis and Synthesis of Linear Servomechanisms*, Albert C. Hall, OSRD 2097, Report to the Services 64, The Massachusetts Institute of Technology, May 1943.
Div. 7-321.1-M3
6. *Antiaircraft Director, T-15-E1*, E. L. Norton, OEMsr-353, Report to the Services 98, Bell Telephone Laboratories, Inc., July 30, 1945.
Div. 7-112.2-M11
7. *Theoretical Calculation on Best Smoothing of Position Data for Gunnery Prediction*, R. S. Phillips and P. R. Weiss, OEMsr-262, AMP Note 11, Report 532, The Massachusetts Institute of Technology, Radiation Laboratory, Feb. 16, 1944.
Div. 14-244.4-M1
AMP-703.4-M11
8. *A Long Range, High-Angle Electrical Antiaircraft Director [Final Report on T-10]*, C. A. Lovell, NDCrc-127, Research Project 2, Division 7 Report to the Services 80, Bell Telephone Laboratories, Inc., June 24, 1944.
Div. 7-112.2-M9
9. *Flight Records of Pitch, Roll, and Yaw*, taken in a variety of bombers at Wright Field, Ohio, Sperry Gyroscope Company, 1942-5.
10. *Design and Performance of Data-Smoothing Network*, R. B. Blackman, OEMsr-262, Report MM-44-110-38, [Bell Telephone Laboratories, Inc.], July 8, 1944.
11. *Computer for Controlling Bombers from the Ground*, E. Lakatos and H. G. Och, OEMsr-262, July 24, 1944.
12. *A Position and Rate Smoothing Circuit for Ground-Controlled Bombing Computers*, R. B. Blackman, OEMsr-262, Report MM-44-110-79, [Bell Telephone Laboratories, Inc.], Aug. 21, 1944.
13. *A Two-Servo Circuit for Smoothing Present Position Coordinates and Rate in Antiaircraft Gun Directors*, R. B. Blackman, Contract W-30-069-ORD-1448, Report MM-44-110-65, [Bell Telephone Laboratories, Inc.], Sept. 27, 1944.
14. *The Theory of Electrical Artificial Lines and Filters*, A. C. Bartlett, John Wiley and Sons, Inc., 1931, p. 28.
15. *Network Analysis and Feedback Amplifier Design*, H. W. Bode, D. Van Nostrand Company, 1945.
- 15a. *Ibid.*, Chapters 7, 8, 13, and 14
- 15b. *Ibid.*, p. 313.
- 15c. *Ibid.*, p. 326.
- 15d. *Ibid.*, p. 301.
- 15e. *Ibid.*, p. 33.
- 15f. *Ibid.*, p. 12.
- 15g. *Ibid.*, p. 78.
- 15h. *Ibid.*, p. 110.
- 15i. *Ibid.*, p. 133.
- 15j. *Ibid.*, Chapter 5.
16. *Fundamental Theory of Servomechanisms*, L. A. MacColl, D. Van Nostrand Company, 1945.
17. *Automatic Control Engineering*, E. S. Smith, McGraw-Hill Book Company, Inc., 1944.
18. *Die Lehre von den Kettenbrücken*, B. G. Teubner, Leipzig, 1913.
19. "Transient Oscillations in Wave Filters," J. R. Carson and O. J. Zobel, *Bell System Technical Journal*, July 1923.
20. "Harmonic Analysis of Irregular Motion," Norbert Wiener, *Journal of Mathematics and Physics*, Vol. 5, 1926, pp. 99-189.
21. "Generalized Harmonic Analysis," Norbert Wiener, *Acta Mathematica*, Stockholm, Vol. 55, 1930, pp. 117-258.
22. "Stochastic Problems in Physics and Astronomy," S. Chandrasekhar, *Review of Modern Physics*, Vol. 15, 1943, pp. 1-89.
23. "Mathematical Analysis of Random Noise," S. O. Rice, *Bell System Technical Journal*, Vol. 23, 1944, pp. 282-332.
- 23a. *Ibid.*, Vol. 24, 1945, pp. 46-156.

[30]

COVER SHEET FOR TECHNICAL MEMORANDA
RESEARCH DEPARTMENT

SUBJECT: The Transient Behavior of a Large Number of Four-Terminal Unilateral Linear Networks Connected in Tandem - Case 20878

ROUTING:

1 - H.W.B.-J.B.F.-H.F.-Case Files	MM- 46-110-49
2 - CASE FILES	DATE April 10, 1946
3 - L.G.Abraham-T.E.Brewer	AUTHOR S C.L. Dolph
4 - C.H.Elmendorf-H.K.Krist	INDEX NO C.E. Shannon
5 - H.S.Black-F.B.Anderson	Index No. W1.416
6 - G.N.Thayer-C.W.Harrison	
7 - R.L.Dietzold	
8 - L.A.MacColl	
9 - B.M.Oliver	
10 - C.L.Dolph	
11- C.E.Shannon	

ABSTRACT

Asymptotic expressions for the transient response of a long chain of four-terminal unilateral linear networks connected in tandem subject to an initial disturbance are developed and classified according to the characteristics of the common transfer ratio. It is shown that a necessary and sufficient condition for the stability of the chain for all n is that the transfer ratio be of the high pass type.

The mathematical results are applied to chains of self-regulating telephone repeaters.

This Copy for

The Transient Behavior of a Large Number of Four-Terminal
Unilateral Linear Networks Connected in Tandem - Case 20878

MM-46-110-49

April 10, 1946

MEMORANDUM FOR FILE

Introduction

The transient response behavior of a long chain of invariable four-terminal networks connected unilaterally in tandem is of primary importance in the design of cross-country wire communication systems, since the successful operation of such equipment depends upon the rapid damping of transients caused by suddenly applied inputs.

While the emphasis in the memorandum will be directed toward coaxial systems consisting of self-regulating repeaters spaced at 3-7 mile intervals and spanning distant points, the results are of a more general nature and would apply, with obvious modifications and corresponding interpretations, to any configuration involving a large number of four-terminal linear invariable networks connected unilaterally in tandem.

It will be shown that there are two fundamentally different types of transient response possible depending upon the gain characteristic of the transfer ratio of the individual four-terminal linear networks comprising the system. The first type of response while satisfactory is difficult to achieve in practice because of the stringent requirements on the gain characteristic of the transfer ratio. The second, a case often encountered in practice, will be shown to be unsatisfactory in general since it leads to build-up and overloading in any physical system comprising a large number of such networks. However, a guiding design principle will be suggested which, it is believed, will enable us to minimize the worst of the effects, and make the successful operation of a system of the type envisaged here possible.

This memorandum is divided into two parts. In the first the problem is defined physically and then formulated mathematically. Following this, the history of the problem is discussed briefly after which the new results are summarized.

Finally, this part concludes with a discussion of their interpretation and implications for the coaxial system. The second part presents the detailed mathematical arguments which led to the new results of part one.

PART I

Statement of the Problem

The analysis in this memorandum is directed toward the understanding of certain anomalous effects which a long chain of self-regulating telephone repeaters may exhibit at its output when the input end of the chain is subject to a transient disturbance (Cf. Figure 1).

The gain settings of the repeaters in such a chain are usually controlled by the level of a pilot frequency somewhere in the communication band and the regulation is designed to compensate for low frequency phenomena (up to approximately one cycle per second) such as the diurnal change in line resistance. The repeaters in the chain are normally absolutely stable devices so that any transient which is presented to the input of any one of them will be evanescent in time at the output of that repeater.

Since transients are not damped out instantaneously even in absolutely stable devices, a transient disturbance at the input to the first repeater in such a chain will be propagated down the chain. It has been experimentally observed that under certain conditions the maximum amplitude of a transient disturbance may increase as the disturbance is propagated from one repeater to the next and in some cases there may be many oscillations of sufficiently large amplitude to render the system inoperative because of prolonged over-loading.

If the entire chain from its input to its output end is considered as a whole, the chain does behave then in many respects like an unstable non-linear device in spite of the fact that each repeater in the chain is absolutely stable.

Since it is obvious that the above type of behavior is at best undesirable in a cross-country link, it is necessary that its cause be thoroughly understood and that all possible steps be taken either to suppress it or, if this is not possible, at least to minimize its effects.

Although it is not reasonable to expect that transient oscillations can be kept from propagating down the line, or that it is possible to isolate the line from all transient disturbances, it is reasonable to seek a means of guaranteeing that the transients that are propagated down the line will never possess amplitudes that exceed the magnitude of the original disturbance or to seek a way to guarantee that the maximum response of the transient oscillations will occur so shortly after the initial disturbance that physical apparatus will be incapable of following or distinguishing it from the unavoidable initial disturbance. A way of guaranteeing the first of these will be discussed at length and a suggestion will be made which it is felt will guarantee the second, although no rigorous proof of this last fact has yet been given.

Fig. 2 represents a schematic drawing of a typical satisfactory type of transient response which might result from a unit step input to the first unit of Fig. 1. Fig. 3, on the other hand, represents a schematic drawing of a typical unsatisfactory type of transient response which could result from the same input to a system of the type of Fig. 1 which had different characteristics. Briefly then, the problem to be discussed is that of determining the relationships between the network characteristics and the transient response for networks of the form of Fig. 1.

Mathematical Formulation of the Problem

A sudden change in level in the pilot frequency before the n-th repeater results in the modulation of this frequency, changing it from its normal form

$$A \sin \omega_c t$$

to

$$A \sin \omega_c t [1 + f(t)]$$

where $f(t)$ represents the modulation introduced by the transient.

After passage through the n-th repeater, this last expression is transformed into

$$A \sin (\omega_c t + \phi) [1 + g(t)],$$

where the repeater and regulator have (possibly) changed the carrier by the addition of the phase angle φ and have modified the original envelope $A[1 + f(t)]$ into $A[1 + g(t)]$.

It is clear that from the standpoint of regulation it is sufficient to limit discussion to the transformation of $f(t)$ into $g(t)$.*

The exact relationship between $f(t)$ and $g(t)$, of course, depends upon the characteristics of the repeater-regulator circuits which are in general non-linear. However, for small signal inputs their behavior may be satisfactorily represented by that obtained from a linear invariable four-terminal network. Thus, the chain of self-regulating repeaters may be replaced, for the purpose of mathematical analysis, by a chain of linear invariable four-terminal networks having a common transfer ratio $y(p)$. Thus, the blocks of Fig. 1, will be idealized as being such linear four terminal networks throughout the analysis.

Because regulation is designed to compensate for low frequency phenomena, certain characteristics that $y(p)$ should possess are known a priori: namely;

- (1) $y(p)$ must represent a high-pass system. That is, $y(p) \rightarrow 1$ as $p \rightarrow \infty$
- (2) $y(0)$ should be zero if, in the terminology of servo theory, there is to be no static error.

In terms of $y(p)$, the design of a self-regulating system reduces to two problems:

(I) Given $y(p)$, to calculate the transient behavior of the chain of self-regulating repeaters,

(II) The design of a system having a $y(p)$ which leads to satisfactory transient behavior.

The rest of the memorandum will be concerned largely with the first of these. The calculations will be carried out in general terms and the different types of possible responses will be described in terms of the characteristics of $y(p)$.

* Transit time between repeaters is neglected throughout this memorandum. More exactly, we choose a different origin of time at each repeater, so that the transit time does not appear explicitly in the formulae.

Mathematically the problem discussed in this memorandum can be formulated as follows: If $y(p)$ represents the common steady-state transfer ratio of the four-terminal linear units shown connected in tandem in Figure 1, the output voltage response of the n -th unit $V_n(t)$ is given by the inverse Laplace integral:

$$(1) \quad V_n(t) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} y(p)^n e^{pt} \hat{V}_0(p) dp$$

where $\hat{V}_0(p)$ represents the spectrum of the input voltage.

For an impulsive input of intensity V_0 applied at time $t = 0$,

$$\hat{V}_0(p) = V_0.$$

For a step function input of height V_0 applied at time $t = 0$,

$$\hat{V}_0(p) = V_0/p.$$

Specifically, this memorandum will be devoted to the study of the behavior of $V_n(t)$ for large values of n .

Four-terminal networks are normally classed as low-, band-, or high-pass depending upon the character of $|y(i\omega)|$. Typical examples of $|y(i\omega)|$ are shown in Figure 4a, in which, following the usual practice, $|y(i\omega)|$ has been normalized to be unity at $\omega = 0$ in the low-pass case; at $\omega = \omega_0$, (the mid-band frequency), in the band-pass case; and at $\omega = \infty$ in the high-pass case.

From the viewpoint of the asymptotic behavior of the system in Figure 1, it is convenient to modify this classification somewhat when speaking of the over-all gain characteristic, $|y(i\omega)|^n$, of the transfer ratio of a system comprised of n units. For sufficiently large n , it is clear that $|y(i\omega)|^n$ would lead to curves of the type shown in Figure 4b corresponding to the low-pass, band-pass and high-pass curves of Figure 4a. Thus, for sufficiently large n , the gain curves B', C', and D' of

Figure 4b are seen to exhibit the type of behavior normally associated with a band-pass characteristic. A' and E', on the other hand, exhibit behavior of the type normally classified as low-pass and high-pass. For these reasons, the terms low-, and high-pass will henceforth be reserved for those gain characteristics which are always less than their values at $\omega = 0$ and $\omega = \infty$, respectively. The term, band-pass, will be used to cover all other cases; namely, those in which $|y(i\omega)|$ possesses one or more maxima at finite frequencies, the values of which exceed the values of $|y(i\omega)|$ at both zero and infinity.

History of the Problem

Several people have considered this problem in the above mathematical form. Before proceeding to a discussion of the results of the general theory, it will be instructive to consider a few illustrative examples of their results.

Let

$$(2)^* \quad y(p) = p/(p+1).$$

The gain characteristic is clearly of the high-pass type and satisfies (1) and (2) of Page 6. If the input voltage is a unit step, then, by the theorem of residues,

$$V_n(t) = \frac{d^{n-1}}{d(t)^{n-1}} \left[\frac{e^{pt} p^{n-1}}{(n-1)!} \right]_{p=-1} = e^{-t} L_{n-1}(t)$$

where $L_{n-1}(t)$ denotes the Laguerre polynomial of degree $(n-2)$. A plot of $V_n(t)$ for $n = 1, 2, \dots, 10$ is shown in Figure 5. It is known that for large n

$$L_n(t) \approx \frac{1}{\sqrt{n}} e^{\frac{t}{2}} (nt)^{-1/4} \cos \left[2(nt)^{1/2} - \frac{\pi}{2} \right]$$

*This example was first treated by L. A. MacColl (MM-39-325-166), 9/1/39 and W. H. Wise (MM-38-343-22), 8/2/38. The above treatment follows that of MacColl.

where \approx is to be interpreted as "asymptotically equal to."
Thus

$$V_{n+1}(t) \approx \frac{1}{\sqrt{\pi}} e^{-\frac{t}{2}} (nt)^{-1/4} \cos \left[2(nt)^{1/2} - \frac{\pi}{2} \right]$$

A plot of the approximate "envelope"

$$\frac{1}{\sqrt{\pi}} e^{-\frac{t}{2}} (nt)^{-1/4}$$

is given for $n = 50, 100, 150, 200,$ and 250 in Figure 6.

The response in this case is seen to be both amplitude and frequency modulated, the "instantaneous frequency" in the sense of frequency modulation theory being given by

$$\omega = \frac{d}{dt} (2(nt)^{1/2}) = \sqrt{\frac{n}{t}}$$

while the envelope of the amplitude modulation is approximately exponential. In particular, the type of behavior found here can be considered satisfactory since there is no tendency for the magnitude of the largest overshoot to increase without limit as the number of repeaters is increased. As will be shown later, this type of behavior is typical of any network having a high-pass characteristic in the generalized sense of that term as it has been defined above.

In MM-40-3500-92 dated 10/14/1940, J. G. Kreer and J. H. Bollman concluded that the appropriate $y(p)$ for a self-regulating repeater employing a directly heated thermistor element in the control device was given by

$$y(p) = \frac{p + a}{p + 1}$$

It should be observed that for $a \neq 0$ this transfer ratio does possess static error. L. A. MacColl in MM-40-130-270 treated this case for $|a| < 1$ and found that the system exhibited essentially the same type of satisfactory behavior as that discussed above.

(2)¹ A slightly more complicated example is given by

$$y(p) = \frac{p(p + \alpha)}{(p + 1)^2}.$$

It is easily seen that for $\alpha < \sqrt{2}$, $|y(i\omega)|$ is a high-pass characteristic in that $|y(i\omega)| < 1$ for all finite ω and $|y(i\omega)| \rightarrow 1$ as $\omega \rightarrow \infty$. On the other hand, if $\alpha > \sqrt{2}$, $|y(i\omega)|$ possesses a maximum greater than 1 at some finite frequency. $|y(i\omega)|$ is illustrated by curve I in Figure 7 for $\alpha = 1.4$ (high-pass) and by Figure 8 for $\alpha = 2$ (band-pass). The response $V_n(t)$ to a unit step function is shown in Figures 9 and 10 for these two cases with $n = 1, 2, \dots, 9$. The character of the response is seen to be of a radically different kind for these two values of α .

For $\alpha = 1.4$ the response is seen to be of the same type as that encountered in the first example. For $\alpha = 2$, on the other hand, it seems to represent an oscillation in which the magnitude of the largest overshoot is increasing without limit as n tends to infinity. Later it will be shown that this is in fact the case and that satisfactory operation is impossible for a large number of repeaters in this case.

From this and other considerations L. A. MacColl conjectured that a necessary and sufficient condition that the response $V_n(t)$ be bounded for all n was that the transfer ration $y(p)$ have no net gain at any frequency. Mathematically expressed, a necessary and sufficient condition that

$$|V_n(t)| \leq M \text{ for all } n,$$

where M is independent of n and t , is that

$$(M) \quad |y(i\omega)| \leq 1 \quad \text{for all real frequencies } \omega.$$

Physically, the condition on $y(i\omega)$ prevents the transfer ratio $|y(i\omega)|^n$ for a system using n units from having a tremendous gain at any particular frequency.

¹This case was also treated by L. A. MacColl, but no memorandum on it was ever written.

In one sense this memorandum could be summarized as a proof of this conjecture. In particular, a direct proof of the necessity of MacColl's condition (M) is given in the second part. The remainder of that part is devoted to an indirect proof of the sufficiency. The argument consists in exhibiting the two types of possible responses; the first being that associated with a $y(p)$ satisfying MacColl's condition and that second that resulting from a $y(p)$ which violates it at one or more frequencies.

Statement of Results

The detailed results of the sufficiency argument are discussed conveniently in terms of the generalized characterization of high-, band-, and low pass $y(p)$'s as given on page 8. The results will be taken up in that order.

High Pass

In terms of the above classification, the class of high pass $y(p)$'s consists of just those functions which satisfy MacColl's condition and are therefore those from which a satisfactory response could be expected. For the $y(p)$'s in this class, it is clear on physical grounds that the maximum contribution to the response $V_n(t)$ of equation (1) will come from the large values of $|\omega|$ since for these values of $|\omega|$, $|y(i\omega)|^n \rightarrow 1$ while for all other values of $|\omega|$, $|y(i\omega)|^n \rightarrow 0$. Using the first three terms of the Laurent expansion of $y(i\omega)$ about $\omega = \infty$, one finds:

$$(5)^* \quad y(i\omega) \approx 1 + \frac{a}{\omega} + \frac{b}{\omega^2},$$

$$(6) \quad |y(i\omega)| \approx \left[1 + \frac{a^2 + 2b}{\omega^2} \right]^{1/2},$$

$$(7) \quad \text{Angle } y(i\omega) \approx \frac{a}{\omega}.$$

* It is assumed that $a > 0$, $b < 0$, and that $2b + a^2 < 0$. These assumptions correspond to a second order maxima at $|\omega| = \infty$ and to a monotonic decreasing phase function for $y(p)$ as $|\omega| \rightarrow \infty$.

If these approximations, which are valid for $|\omega|$ sufficiently large, are introduced into equation (1), it can be shown that the principal contribution to $V_n(t)$ for a unit step input is given by:

$$V_n(t) \approx (\pi)^{-1/2} (nat)^{-1/4} \exp \left\{ \left(\frac{a^2 + 2b}{2a} \right) t \right\} \cos \left(2\sqrt{nat} \frac{-\pi}{4} \right).$$

This, with a suitable interpretation of the constants a and b is seen to be of the same general form as the response obtained by MacColl for $y(p) = p/(p + 1)$ as given by equation (3). Just as in that example the response is both frequency and amplitude modulated. The instantaneous frequency of oscillation is again given by

$$\omega = \frac{d}{dt} \left(2\sqrt{nat} \frac{-\pi}{4} \right) = \sqrt{\frac{n a}{t}}.$$

The gain for

$$y(p) = \frac{p(p + 0.5)}{(p + 1)^2}$$

is shown on curve I of Figure 11. Curve II of this figure represents $|y(i\omega)|^{100}$ for this $y(p)$. For this example and $n = 100$, the true gain $|y(i\omega)|^{100}$ and the gain approximation resulting from equation (6) are indistinguishable on the scale of Figure 11.

The corresponding phase characteristic for $y(p)^{100}$ is plotted on Figure 12 where, for reasons which will appear in Part II, the actual frequency has been replaced by

$$\omega' = \frac{\omega}{\sqrt{n}}.$$

Again, on the scale of Figure 12 the actual phase is indistinguishable from the approximation resulting from equation (7). Figs. 7 and 13 present the same information for

$$y(p) = \frac{p(p + 1.4)}{(p + 1)^2} \quad \text{and } n = 100.$$

Again the agreement between the actual phase and the approximation is excellent. However, there is a considerable error in the gain approximation for small $|\omega|$. This large error is unquestionably due to the fact that the value $\alpha = 1.4$ is near the critical value $\alpha = \sqrt{2}$ at which the characteristic changes from high-pass to band-pass.

Agreement with the above asymptotic formula can of course be obtained by increasing n sufficiently. Alternately, for $n = 100$, a better approximation to the gain can be obtained by writing

$$y(i\omega) \approx 1 + \frac{a}{\omega} + \frac{b}{\omega^2} + \frac{c}{\omega^3}$$

and

$$|y(i\omega)| \approx \left[1 + \frac{2b + a^2}{\omega^2} + \frac{2d + b^2 + 2ac}{\omega^4} \right]^{1/2}$$

This approximation leads to a curve which is indistinguishable from that of $|y(i\omega)|_{100}$ in Figure 7. With this approximation, one finds the following expression for $V_n(t)$ when the input is a unit step function

$$\dot{V}_n(t) \approx (\pi)^{-1/2} (\text{nat})^{-1/4} \cos \left(2\sqrt{\text{nat}} \frac{-\pi}{4} \right) \exp \left(\frac{(a^2 + 2b)t}{2a} \right)$$

$$\left(1 + \frac{(2d + b^2 + 2ac)t^2}{2a^2 n} \right)$$

This expression is seen to approach that given by equation (8) as $n \rightarrow \infty$. Thus one can conclude that the response will always be satisfactory if $y(p)$ belongs to the class of high-pass characteristics.

Band-Pass Case

MacColl's condition is clearly violated whenever $|y(i\omega)|$ has one or more relative maxima greater than 1 at finite frequencies. For simplicity the case where $|y(i\omega)|$ has only one such

maxima at $\omega = \omega_0$ will be treated first. It will furthermore be assumed that this maximum is of the second order; i.e.

$$\frac{d^2}{d\omega^2} |y(i\omega)|_{\omega = \omega_0} \neq 0.$$

Under these conditions, it is physically clear that the maximum contribution to the response $V_n(t)$ as given by equation (1) will be due to those frequencies near ω_0 , at which $|y(i\omega)|$ possesses its maximum, since as n increases this region becomes increasingly more important than all the rest. It is also clear that the time of maximum response will be given by the delay time experienced by the frequency ω_0 in passing thru the network. This is known to be given by $t_0 = -n B'(\omega_0)$ where $B'(\omega_0)$ denotes the slope of the phase characteristic $B(\omega)$ in the expression

$$(10) \quad y(i\omega) = A(\omega) \exp(iB(\omega)).$$

If $A(\omega)$ and $B(\omega)$ are expanded in a Taylor's series about $\omega = \omega_0$ and terms up to the second order retained, it can be shown that the response to a unit impulse function is given by

$$(11) \quad V_n(t) \approx \frac{A(\omega_0)^n}{\sqrt{2n}} G(\omega_0) \exp\left(-\frac{(t-t_0)^2 H(\omega_0)}{2n}\right) \cos\left[\omega_0 t + nB(\omega_0) + \varphi(\omega_0)\right]$$

where

$$G(\omega_0) = \pi^{-3/2} \left\{ \left[\frac{A''(\omega_0)}{A(\omega_0)} \right]^2 + [B''(\omega_0)]^2 \right\}^{-1/4}$$

$$H(\omega_0) = \frac{A''(\omega_0)}{A(\omega_0) \left\{ \left[\frac{A''(\omega_0)}{A(\omega_0)} \right]^2 + [B''(\omega_0)]^2 \right\}} > 0$$

$$\phi_o(\omega_o) = \arctan\left(\frac{B''(\omega_o) A(\omega_o)}{2A'(\omega_o)}\right)$$

$$t_o = -nB(\omega_o).$$

Thus $V_n(t)$ can be interpreted as an amplitude modulated wave with an envelope proportional to the Gauss error curve

$$\exp\left(\frac{-(t-t_o)^2}{2n}\right) H(\omega_o)$$

with a standard deviation given by

$$\sigma = \left\{ n \frac{A(\omega_o)}{A'(\omega_o)} \left[\frac{\{A'(\omega_o)\}^2}{\{A(\omega_o)\}^2} + \{B'(\omega_o)\}^2 \right] \right\}^{1/2}$$

The standard deviation σ is of course a convenient measure of the duration of the disturbance. The maximum response occurs for time $t_o = -n B'(\omega_o)$ at which time the amplitude is proportional to

$$\frac{A(\omega_o)^n}{\sqrt{2n}}$$

Thus if $A(\omega_o) > 1$, the maximum response will represent a value which is very large compared with unity, the magnitude of the original disturbance, if n is large. This would force any system involving vacuum tubes to overload if n were sufficiently large.

These properties are summarized in Figures (14) and (15). Figure (14) is a plot of the response for values of t near t_o for a few values of n for the example given by equation (4) where $\alpha = 2$. Figure (15) is a plot of the maximum response for a few values of n for different values of the parameter α .

It should be remarked that the above approximation to the gain which was obtained by keeping only the first two terms

of the expansion of $A(\omega)$ about $\omega = \omega_0$ could only be expected to be a reasonable one for fairly large values of n , since it represents a usually unsymmetric gain characteristic by a symmetric function. A better or second approximation can be obtained by using three terms of the Taylor's expansion instead of two. Just as in the high pass case, the retention of this extra term gives rise to a second term in the expression for $V_n(t)$ but it does not fundamentally alter the characteristics of the response since the correction term vanishes for $t = t_0$, at which time the response is still a maximum, with the same amplitude as before. Its only effect is to take cognizance of the unsymmetrical character of the gain characteristic $A(\omega)$ and to change the resulting response envelope to an unsymmetrical one. Of course, it also modifies the phase of the oscillation inside the envelope in a complicated way without changing the fundamental frequency of oscillation.

For these reasons and because of the complexity of the resulting expression, it will not be written down here explicitly although the explicit approximation to the gain $A(\omega)$ will be discussed in Part II.

The two approximations to the gain are illustrated for equation (4) with $\alpha = 2$ in Figure 16 for $n = 100$. In this case

$$A(\omega) = \frac{|\omega| \sqrt{\omega^2 + 4}}{\omega^2 + 1}.$$

As can be seen from the figure, the second approximation does in fact represent $A(\omega)$ over the significant range of frequencies near ω_0 from which it can be concluded that the response will be unsatisfactory. Figure (14), previously referred to, furnishes a picture of the envelope response as obtained from the first approximation.

In the event that $A(\omega)$ takes on its maximum value at more than one place in the finite frequency range, it is clear that the above results can be generalized as follows:

Let $V_{ni}(t)$ be the response of the form given by equation (11) due to a maximum at $\omega = \omega_i$. Let the time of maximum response

from this maximum be denoted by $t_i = -nB'(\omega_i)$. Then the total response is clearly given by the expression

$$V_n(t) \approx \sum_{i=1}^k V_{ni}(t),$$

if there are k relative maxima. Unless the values of $A(\omega)$ at the points $\omega = \omega_i$ are nearly the same, it is also clear that only those terms of the above sum which correspond to the largest maxima of $A(\omega)$ will be of significance.

The band-pass case is also discussed briefly for unit step inputs in Part II.

Low Pass Case

Since the low-pass case differs from the band pass case only in that $A(\omega)$ has its maximum for $\omega = 0$ instead of at $\omega = \omega_0 \neq 0$ the results of the two are very similar. The results in the low-pass case are simpler because it will be recalled that $B(\omega)$ (as defined by equation 10) is an odd function of ω for any physical network. This forces both $B(0)$ and $B''(0)$ to be zero so that for an impulsive input one obtains the simple formula:

$$(12) \quad V_n(t) \approx \frac{An(0)}{\sqrt{2n}} \left\{ \pi^{-3/2} \left[-\frac{A''(0)}{A(0)} \right]^{-1/2} \right\} \exp \left\{ \frac{(t-t_0)^2 A(0)}{2n A''(0)} \right\}$$

This result corresponds to the well-known formula from transmission line theory for non-distortionless lines.

Remarks

From the practical viewpoint the above results have the following implications for communications systems such as a cross-country coaxial telephone system employing self-regulation repeaters spaced at intervals of a few miles.

(1) If the transfer characteristic of each individual network is of the high-pass type (in the sense in which this term has been used above) then the transient response will never exceed the initial value of the disturbing input voltage and it will be damped out so that the operation of the communication system would generally be considered satisfactory.

(2) If the network is not of the high-pass type, the usual practical case, and there is any net gain in the system, which is peaked at ω_0 then for even a small number of units the response will exceed the initial input at the time given by

$$t_0 = -nB'(\omega_0)$$

where

$$A'(\omega_0) = 0$$

and if the number of units is sufficiently large the output from the n -th unit will be large enough to cause severe overloading.

At first glance these implications are not promising and seem to indicate that the operation of a cross-country system involving several hundred repeaters and regulators would be extremely difficult, since the only satisfactory characteristic is difficult to attain in practice. However, practically the ideal characteristic which is high pass can be approached in the sense that the peaked frequency can be made very large. Thus the maximum response may occur so soon after the initial disturbance that the physical system would not be able to follow it or to distinguish it from the initial disturbance which in many cases would be large enough to cause momentary overloading of the system.

Moreover, it is an experimental fact that in the design of feedback regulator characteristic forcing the peaked frequency higher reduces the size of the peak which in turn will permit the use of a larger number of regulators in the system.

If this is done, the time of maximum response, $t_0 = nB'(\omega_0)$, will be small since $B'(\omega)$ in general is small for large ω . Assuming that the effects of the maximum response have been treated in this way, it is natural to inquire into the type of response which will result for finite values of $t > t_0$.

If one examines the gain characteristic curve of the type shown in Figure (7), it is clear that for frequencies less than some frequency ω , slightly less than the peak frequency ω_0 ,

the shape is fundamentally like that of the high-pass case. Remembering that the phase delay of a frequency through a linear network is given by the slope of phase characteristic at that frequency, it is clear that the response for values of t greater than t_0 , the time of maximum response, will come from the frequencies less than ω_0 , since the phase slope characteristic is large for small frequencies and small for large frequencies. Now if it is assumed that the phase characteristic $nB(\omega)$ is a monotonic decreasing function of ω , it is clear that the function $(nB(\omega) + \omega t)$ will always be stationary at an arbitrary frequency ω , provided that t is given a suitable corresponding value. Thus, it is reasonable to expect that the response for $t \gg t_0^*$ will exhibit the same type of character as that obtained in the high-pass case discussed above. This, it will be recalled, is both frequency and amplitude modulated with an envelope which decreases approximately exponentially. Thus, under these circumstances it seems reasonable to suppose that satisfactory operation of the communication link could be obtained.

To recapitulate, the most practical design for any system of the type envisaged in Figure 1, from the viewpoint of satisfactory transient response involves approaching the high-pass characteristic as closely as possible by making the gain characteristic of the transfer ratio peak at as high a frequency as is practicable and by keeping the phase slope characteristic monotonic for all smaller frequencies.

PART II

Mathematical Discussion

Theorem I. A necessary condition that the response $V_n(t)$ from a chain of n -four terminal linear invariable networks subject to a unit step input function have a common finite bound for all n is that the transfer ratio $y(p)$ satisfy the relation

$$(M) \quad |y(i\omega)| \leq 1 \text{ for all real values of } \omega.$$

* A different type of expansion, valid for any fixed t or $n \rightarrow \infty$ is discussed at the end of Part II.

Proof: By hypothesis

$$|v_n(t)| \leq M \text{ for all } n \text{ where } M \text{ is independent of } n \text{ and } t$$

By definition:

$$\hat{v}_n(p) = \int_0^{\infty} e^{-pt} v_n(t) dt$$

$$y(p)^n = \frac{\hat{v}_n(p)}{\hat{v}_0(p)} = p \hat{v}_n(p)$$

so that

$$|y(p)|^n = |p| \left| \int_0^{\infty} e^{-pt} v_n(t) dt \right|$$

$$\leq |p| \int_0^{\infty} |e^{-pt}| |v_n(t)| dt$$

$$\leq |p| M \int_0^{\infty} |e^{-pt}| dt.$$

If $p = c + i\omega$ and if $c > 0$, then

$$|y(p)|^n \leq \frac{|\sqrt{c^2 + \omega^2}| M}{|c|}$$

so that

$$\log |y(p)| \leq \frac{1}{n} \log \frac{M|\sqrt{c^2 + \omega^2}|}{|c|}$$

Thus, in the limit as $n \rightarrow \infty$, it follows that for any p with a positive real part

$$\log |y(p)| \leq 0$$

and hence

$$|y(p)| \leq 1$$

Since this relation holds everywhere in the right-hand half plane, it follows from simple continuity considerations that the maximum of $|y(i\omega)|$, never exceeds 1. Thus

$$|y(i\omega)| \leq 1$$

as was to be shown.

The remaining discussion will be devoted to the characterization of the different types of possible responses and will, at the same time, furnish an indirect proof of the fact that the condition (M) on $y(p)$ is also sufficient.

High Pass Case - Unit Step Input

If the networks comprising the system shown in Figure 1 possess a transfer ratio having a high pass gain characteristic in the sense defined above, and if one writes

$$y(i\omega) = A(\omega) e^{iB(\omega)}$$

then the gain function $A(\omega)$ satisfies the two conditions

(A) $A(\omega) < 1$ for all finite frequencies ω .

(B) $\lim_{\omega \rightarrow \infty} A(\omega) = 1$

Under these conditions it is clear that, for sufficiently large n , the main contributions to $V_n(t)$ will be due to the high values of $|\omega|$. For convenience, $V_n(t)$ is written here in slightly different form

$$V_n(t) = \operatorname{Re} \left\{ \frac{1}{\pi} \int_0^\infty A(\omega)^n e^{i[nB(\omega) + \omega t - \frac{\pi}{2}]} \frac{d\omega}{\omega} \right\}$$

For large values of $|\omega|$, all physical transfer ratios $y(i\omega)$ of interest to us here can be represented by an expansion of the form*

$$(13) \quad y(i\omega) = \left(1 + \frac{a_1}{\omega} + \frac{b}{\omega^2} + \frac{c_1}{\omega^3} + \frac{d}{\omega^4} + \dots \right)$$

We shall confine our attention to the ordinary case, in which $a > 0$, $b < 0$ and $2b + a^2 < 0$. For large values of $|\omega|$, we now have

$$(14) \quad A(\omega) = \left(\left[1 + \frac{b}{\omega^2} + \frac{d}{\omega^4} + \dots \right]^2 + \left[\frac{a}{\omega} + \frac{c}{\omega^3} + \dots \right]^2 \right)^{1/2}$$

$$(15) \quad B(\omega) = \arctan \frac{\frac{a}{\omega} + \frac{c}{\omega^3} + \dots}{1 + \frac{b}{\omega^2} + \frac{d}{\omega^4} + \dots}$$

It is clear that, for $|\omega|$ sufficiently large, the leading terms of these expressions will furnish adequate approximations to $A(\omega)$ and $B(\omega)$. These are:

$$(16) \quad A(\omega) = \left[1 + \frac{a^2 + 2b}{\omega^2} \right]^{1/2},$$

$$(17) \quad B(\omega) = \frac{a}{\omega}.$$

Let ω_0 be the frequency defined by the condition that these approximation are accurate to within the arbitrarily chosen permissible error ϵ for values of ω such that $\omega \geq \omega_0$. Then we can write

* In the usual case $y(p)$ is a rational function, so that this expansion can be readily obtained.

$$\begin{aligned}
 v_n(t) &= \frac{1}{\pi} \operatorname{Re} \left(\int_0^{\omega_0} A(\omega)^n e^{i[nB(\omega) + \omega t - \frac{\pi}{2}]} \frac{d\omega}{\omega} \right. \\
 &\quad \left. + \int_{\omega_0}^{\infty} \left[1 + \frac{(a^2 + 2b)}{\omega^2} \right]^{n/2} e^{i\left[\frac{na}{\omega} + \omega t - \frac{\pi}{2}\right]} \frac{d\omega}{\omega} \right) \\
 &= \frac{1}{\pi} \operatorname{Re} (I_1 + I_2).
 \end{aligned}$$

It is clear that

$$|I_1| \leq \int_0^{\omega_0} \frac{[A(\omega)]^n}{|\omega|} d\omega.$$

Since $[A(\omega)]^n \rightarrow 0$ for each ω in the finite range $0 \leq \omega \leq \omega_0$, it is clear that $|I_1|$ can be made negligibly small by taking n sufficiently large. Introducing the new variable v defined by the relation

$$3) \quad v = \frac{\omega}{\sqrt{\frac{na}{t}}}$$

I_2 can be written as

$$I_2 = \int_{v_0}^{\infty} \left[1 + \frac{(a^2 + 2b)t}{na v^2} \right]^{n/2} e^{i\sqrt{nat} \left(\frac{1}{v} + v \right) - \frac{\pi}{2}} i \frac{dv}{v}.$$

Letting

$$\gamma = \frac{(a^2 + 2b)t}{av^2}$$

and using the binominal expansion, one has

$$\begin{aligned} \left[1 + \frac{(a^2 + 2b)t}{n a v^2} \right]^{n/2} &= \left[1 + \frac{\gamma/2}{n/2} \right]^{n/2} = 1 + \frac{n}{2} \left(\frac{\gamma/2}{n/2} \right) + \\ &+ \frac{\frac{n}{2} \left(\frac{n}{2} - 1 \right)}{2} \left(\frac{\gamma/2}{n/2} \right)^2 + \dots \\ &= 1 + \frac{\gamma}{2} + \frac{1}{2} \left(1 - \frac{1}{n/2} \right) \left(\frac{\gamma}{2} \right)^2 + \dots \\ &= e^{\gamma/2} + \text{terms in } 1/n. \end{aligned}$$

Thus, for sufficiently large n , I_2 becomes, approximately

$$I_2 = e^{-\frac{\pi}{2} i} \int_{v_0}^{\infty} e^{\frac{(a^2 + 2b)t}{2 a v^2}} e^{i \sqrt{n a t} \left(\frac{1}{v} + v \right)} \frac{dv}{v}.$$

In this form the principle of stationary phase can be applied to I_2 (Cf. Appendix I); for the amplitude factor

$$\frac{e^{\frac{(a^2 + 2b)t}{2 a v^2}}}{v}$$

is independent of n , while the phase function (in the notation of the appendix)

$$\varphi(v) = \left(\frac{1}{v} + v \right)$$

is monotonic in the range of integration on each side of the stationary point ($v = 1$) where

$$\varphi'(v) = 0.$$

Physically speaking the form of equation (18) suggest the interpretation of $V_n(t)$ as the sum of an infinite number of complex waves whose amplitudes are slowly varying function of v and whose complex phases are rapidly varying functions of v . Under this interpretation it is physically reasonable to expect that wave interference will occur everywhere except near $v = 1$ where the phase function given by equation (19) is stationary. This is the principal of stationary phase. It remains to evaluate the principal contribution to I_2 for values of v near 1. Replacing $\varphi(v)$ by the first three terms of its Taylor's series about $v = 1$,

$$\varphi(v) = \varphi(1) + 0 + \frac{\varphi''(1)(v-1)^2}{2!} = 2 + (v-1)^2$$

the main contribution to I_2 is given by

$$I_2 \approx e^{i[2\sqrt{nat} - \frac{\pi}{2}]} \int_{1-\eta}^{1+\eta} \frac{(a^2 + 2b)t}{2av^2} e^{i\sqrt{nat}} (v-1)^2 dv,$$

In the interval $(1 - \eta, 1 + \eta)$, the amplitude factor

$$\frac{1}{v} \exp [(a^2 + 2b)t/2av^2]$$

is substantially constant and may be removed from under the integral sign and evaluated at $v = 1$. By the reasoning of Appendix I, the contributions to the remaining integral are not appreciably affected if the limits are changed to $(-\infty, \infty)$ respectively. Letting

$$\xi = v - 1$$

we can then write I_2 in the form

$$I_2 \approx \exp \left\{ \frac{(a^2 + 2b)t}{2a} \right\} \exp \left[i 2\sqrt{nat} - i \frac{\pi}{2} \right] \int_{-\infty}^{\infty} e^{i\sqrt{nat}} \xi^2 d\xi$$

By the known properties of Fresnel integrals

$$\int_{-\infty}^{\infty} e^{im\xi^2} d\xi = \sqrt{\frac{\pi}{m}} e^{\frac{\pi i}{4}},$$

and hence

$$I_2 \cong \exp \left\{ \frac{(a^2 + 2b)t}{2a} \right\} \exp [i2\sqrt{nat} - \frac{\pi}{4}i] \cdot \pi^{-1/2} (nat)^{-1/4}$$

Taking the real part and dividing by π , the asymptotic expression for $V_n(t)$ is therefore given by:

$$(20) \quad V_n(t) \cong \pi^{-1/2} (nat)^{-1/4} \exp \left\{ \frac{(a^2 + 2b)t}{2a} \right\} \cos (2\sqrt{nat} - \frac{\pi}{4})$$

which is equation (8) of Part I.

A more accurate approximation to the gain $A(\omega)^n$ is given by

$$A(\omega) = \left[1 + \frac{2b + a^2}{\omega^2} + \frac{2d + b^2 + 2ac}{\omega^4} \right]^{1/2}$$

where the first three terms of equation (13) have been retained. From this it follows that:

$$\begin{aligned} A(\omega)^n &\cong \exp \frac{n}{2} \left\{ \frac{2b + a^2}{\omega^2} + \frac{2d + b^2 + 2ac}{\omega^4} \right\} \\ &= \exp \left\{ \frac{n}{2} \frac{(2b + a^2)}{\omega^2} \right\} \exp \left\{ \frac{n}{2} \frac{(2d + b^2 + 2ac)}{\omega^4} \right\} \end{aligned}$$

from which it follows that the second approximation is obtained by multiplying the first by the factor

$$\exp \left\{ \frac{n}{2} \frac{(2d + b^2 + 2ac)}{\omega^4} \right\}$$

If the frequency transformation $v = \frac{\omega}{\sqrt{na}} \sqrt{t}$ is now made the first factor will as before be independent of n . Over the range of integration where the integral is significant, their product can be removed from under the integral sign giving

$$V_n(t) \approx (\pi)^{-1/2} (nat)^{-1/4} \cos (2\sqrt{nat} - \frac{\pi}{4})$$

$$\exp \left[\frac{(a^2 + 2b)t}{2a} \right] \left[\exp \frac{(2d + b^2 + 2ac)t^2}{2a^2 n} \right]$$

$$\approx (\pi)^{-1/2} (nat)^{-1/4} \cos (2\sqrt{nat} - \frac{\pi}{2})$$

$$e^{\frac{(a^2 + 2b)t}{2a}} \left[1 + \frac{(2d + b^2 + 2ac)t^2}{2a^2 n} + \dots \right]$$

which is the equation (9) of Part I.

Band Pass Case - Impulsive Input

For simplicity let it be assumed that the gain characteristic $A(\omega)$ has only one absolute maximum at $\omega = \omega_0$ on the positive frequency range and that this is a second order maximum.

The response $V_n(t)$ can always be written in the form

$$V_n(t) = \frac{A(\omega_0)^n}{\pi} \operatorname{Re} \left\{ \int_0^\infty e^{n \log \frac{A(\omega)}{A(\omega_0)} + i n B(\omega) + i \omega t} d\omega \right\}.$$

In this form, $V_n(t)$ can again be interpreted as being proportional to the sum of an infinite number of complex waves of amplitude

$$\left[\frac{A(\omega)}{A(\omega_0)} \right]^n$$

with varying complex phase* given by

$$\varphi(\omega, t, n) = nB(\omega) + \omega t.$$

With this interpretation it is clear that the maximum contribution to $V_n(t)$ will be given by those frequencies in the neighborhood of ω_0 , where ω_0 satisfies $A'(\omega) = 0$ and at values of the time t near t_0 at which the phase function, $\varphi(\omega, t, n)$ is stationary for the maximum frequency ω_0 . Thus t_0 is given by

$$t_0 = -nB'(\omega_0).$$

Since

$$A(\omega_0) \neq 0 \quad \text{and} \quad A'(\omega_0) = 0$$

*"Phase" as used here differs from the way it is normally used in engineering.

one can write for a suitable small neighborhood of ω_0

$$(21) \quad \log \frac{A(\omega)}{A(\omega_0)} = \frac{A''(\omega_0)}{2A(\omega_0)} (\omega - \omega_0)^2 + \frac{A'''(\omega_0)}{6A(\omega_0)} (\omega - \omega_0)^3 + \dots$$

If we retain only the first term of this expansion, then for a suitably restricted neighborhood of ω_0 , one has

$$(22) \quad \left[\frac{A(\omega)}{A(\omega_0)} \right]^n = e^{n \log \frac{A(\omega)}{A(\omega_0)}} \approx e^{\frac{nA''(\omega_0)}{2A(\omega_0)} (\omega - \omega_0)^2}.$$

Similarly, for ω sufficiently near ω_0

$$(23) \quad B(\omega) = B(\omega_0) + B'(\omega_0)(\omega - \omega_0) + \frac{B''(\omega_0)}{2} (\omega - \omega_0)^2.$$

Henceforth for simplicity, we shall write

$$A = A(\omega_0), \quad A'' = A''(\omega_0), \quad B = B(\omega_0), \quad B' = B'(\omega_0),$$

$$B'' = B''(\omega_0)$$

If these approximations are valid in the neighborhood, $(\omega_0 - \Delta, \omega_0 + \Delta)$ it follows that

$$\begin{aligned} V_n(t) = \frac{1}{\pi} \operatorname{Re} \left\{ \left[\int_0^{\omega_0 - \Delta} + \int_{\omega_0 + \Delta}^{\infty} \right] A(\omega)^n e^{i[nB(\omega) + \omega t]} d\omega \right. \\ \left. + A^n \int_{\omega_0 - \Delta}^{\omega_0 + \Delta} \exp \left[\frac{nA''}{2A} (\omega - \omega_0)^2 + i[nB + nB'(\omega - \omega_0) \right. \right. \\ \left. \left. + \frac{nB''}{2} (\omega - \omega_0)^2 + \omega t \right] d\omega \right\}. \end{aligned}$$

Since $[A(\omega)]^n \rightarrow 0$ as $n \rightarrow \infty$, except near $\omega = \omega_0$, it follows as before that the sum of the bracketed integrals can be made negligibly small in comparison with the remaining one if n is taken sufficiently large. Recalling that

$$t_0 = -nB'(\omega_0)$$

the remaining integral can be written as

$$V_n(t) = \frac{1}{\pi} \operatorname{Re} \left\{ A^n e^{i[nB + \omega t]} \int_{\omega_0 - \Delta}^{\omega_0 + \Delta} \exp \left[\frac{n A''}{2A} (\omega - \omega_0)^2 + i(t - t_0)(\omega - \omega_0) + \frac{inB''}{2} (\omega - \omega_0)^2 \right] d\omega \right\}$$

Again the finite limits of integration can be replaced by $-\infty$ and ∞ since, for large n ,

$$e^{\frac{n}{2} \frac{A''}{A} (\omega - \omega_0)^2}$$

will be small except in the immediate neighborhood of ω_0 .

If one sets

$$\rho = -n \left(\frac{A''}{2A} + cB'' \right)$$

$$\rho^2 = i^2(\omega - \omega_0)^2 ; g = t - t_0$$

then the remaining integral can be recognized as pair No. 710.0 of the Campbell and Foster Tables.

Then one finds

$$V_n(t) \approx \frac{1}{2\pi^{3/2}} \operatorname{Re} \left\{ \frac{A^n \exp[inB + i\omega_0 t]}{\sqrt{\rho}} \exp \left[\frac{-(t-t_0)^2}{4\rho} \right] \right\}$$

The result is equivalent to that given by equation (11) of part I. If $A(\omega_0)$ is greater than 1, it is thus seen that the response will have a maximum value that builds up very rapidly as n increases and would eventually force any system involving vacuum tubes to overload.

It should be remarked that the above approximation to the gain could only be expected to be a reasonable one for fairly large values of n , since it represents a usually unsymmetric gain characteristic by a symmetric function. A better or second approximation can be obtained by keeping the second term of the expansion of the logarithm in (21), and then taking the first term of the expansion of

$$e^{\frac{nA'''}{6A} (\omega - \omega_0)^3}.$$

This yields

$$[A(\omega)]^n \approx A^n \left\{ e^{\frac{nA''}{2A} (\omega - \omega_0)^2} \left[1 + \frac{nA'''}{6A} (\omega - \omega_0)^3 \right] \right\}.$$

The addition of the second term in the above expression gives rise to an additional term in $V_n(t)$, provided that the same phase approximation (23) is retained. The resulting $V_n(t)$ is similar to (11) but the new envelope consists of the old envelope plus $nA'''/6A$ times the third derivative of the old envelope. The modulated frequency remains the same but the phase is changed in a complicated manner. (Compare pair 710.3 of the Campbell and Foster tables).

Unit Step Input

In this case one can write

$$V_n(t) = \frac{1}{\pi} \operatorname{Re} \left\{ \int_0^{\infty} \frac{A(\omega)^n}{\omega} e^{i[nB(\omega) + \omega t \frac{\pi}{2}]} d\omega \right\}.$$

As before the only significant frequencies are in the neighborhood of $\omega = \omega_0$ and near this point the $\frac{1}{\omega}$ in the denominator can be taken out of the integral as $1/\omega_0$ provided $\omega_0 \neq 0$. Thus the result will be same as for the impulsive input apart from the factor $1/\omega_0$ if one makes $nB(\omega) - \pi/2$ correspond to $nB(\omega)$ in (11).

Low-Pass Case

It is clear that the analysis for this case in which the equation $A'(\omega) = 0$ is satisfied for $\omega = 0$ can be carried through in exactly the same manner as the band-pass case treated previously. The resulting answer is capable of simplification, however, if it is recalled that $B(\omega)$ for any physical network is an odd function of ω . This forces both $B(0)$ and $B''(0)$ to be zero. The resulting formulae then become

a) Impulsive Input

$$V_n(t) = \frac{A(0)^n e^{-\frac{(t-t_0)^2 A(0)}{2n A'''(0)}}}{\pi^{3/2} \sqrt{\frac{2n A'''(0)}{A(0)}}}.$$

b) Unit Step Input

(24)

$$V_n(t) = \frac{A(0)^n}{\pi^{3/2} \sqrt{\frac{2n A'''(0)}{A(0)}}} \int_0^t \exp \left\{ \frac{-(t-t_0)^2 A(0)}{2n A'''(0)} \right\} dt.$$

This last expression involves an integral since it is necessary to eliminate the pole at zero where $A(\omega)$ has its maximum. This can be done by differentiating $V_n(t)$ with respect to t , finding the asymptotic formula for $V'_n(t)$ as before and then integrating to obtain (24).

Hamy's Expansions in the Band-Pass Case

The type of asymptotic expansions so far given for the band-pass case were explicitly designed to represent $V_n(t)$ in the neighborhood of $t = t_0$ where $V_n(t)$ is a maximum. They could in no sense be considered the true asymptotic expansions for values $t \ll t_0$ or $t \gg t_0$. In particular their derivation depended upon the fact that the time of maximum response was related to the number of four terminal networks by means of the equation

$$t_0 = -nB'(\omega_0),$$

so that as $n \rightarrow \infty$, $t_0 \rightarrow \infty$.

Other types of expansion are clearly possible. Two obvious alternatives are:

- (1) Those valid for fixed n as $t \rightarrow \infty$;
- (2) Those valid for fixed t as $n \rightarrow \infty$.

The first of these will not be considered here since they are of little interest as all of the four terminal networks have been assumed to be absolutely stable. The interested reader is referred to the book by Doetsch on Laplace Transformations for expansions of this type.

Since the second type of expansion is of interest here and is not to be found in most of the standard reference works it will be discussed here briefly.

In a classic paper, M. Hamy* derived general expansions of this type for complex integrals of the form

$$\int f(z) \varphi^n(z) dz$$

*Journal de Mathematique, vol. 4, 6th series, 1908, page 203.

under a variety of hypotheses on $f(z)$ and $\varphi(z)$. These conditions include the case where $\varphi(z)$ has a saddle point given by the solution of $\varphi'(z) = 0$ and the result of this case is a generalization of the often-used theorem of Fowler which one finds in his book on statistical mechanics under the title of the saddle point method.

More to the point, they also include the case where $\varphi(z)$ has one or more maxima on the path of integration at which $\varphi'(z) = 0$ provided that $f(z)$ admits a Taylor series expansion about these points. In particular, then, if one considers t as a fixed parameter they apply to the integral of equation (1), with $c = 0$ and $\varphi(z) = y(p)$; $f(z) = e^{pt} \tilde{V}_0(p)$.

In terms of our notation, one finds that:

(a) for an impulsive input with gain maxima at $\omega = \omega_0$

$$V_n(t) \cong \frac{2A^n(\omega_0)}{nB'(\omega_0)} \cos [\omega_0 t + n B(\omega_0)] + \text{term in } \frac{1}{n^2}.$$

(b) for a unit step input with gain maxima at $\omega = \omega_0 \neq 0$.

$$V_n(t) \cong \frac{2A^n(\omega_0)}{nB'(\omega_0)\omega_0} \cos [\omega_0 t + nB(\omega_0) - \frac{\pi}{2}] + \text{terms in } \frac{1}{n^2}.$$

It is interesting to note that these formula indicate a dependence upon $1/n$ instead of $1/\sqrt{n}$ as in the case of the previous expansion. These formulae can be thought of as representing the response in the band-pass case for any fixed t , $t \ll t_0$.

Appendix I

Certain remarks of Auerl Winter* on the justification of the principle of stationary phase are pertinent enough to the above discussion to bear repetition here. In order for the integral

$$(25) \quad S = \int_a^b f(x) e^{i\rho\varphi(x)} dx$$

to be asymptotically represented as $\rho \rightarrow \infty$, by the formula (Cf. Lamb, Hydrodynamics p 395)

$$(26) \quad S \approx \frac{\sqrt{\pi f(\alpha)}}{\sqrt{\frac{1}{2}\rho|\varphi''(\alpha)|}} e^{i[\rho\varphi(\alpha) \pm \frac{1}{4}\pi]}$$

where $\varphi'(\alpha) = 0$ and where the upper or lower sign is to be taken according as $\varphi''(\alpha)$ is positive or negative, it is evident that two things are sufficient.

- (1) The contribution to the integral outside a small interval around the stationary value α of $\varphi(\alpha)$ must decrease more rapidly as a function of ρ than the one obtained in the neighborhood of α ;
- (2) The asymptotic formula given above must adequately represent the behavior of the contribution to the integral from the neighborhood of the stationary value α .

Now, if, on any closed interval I , $\varphi'(x)$ is continuous and has no zeros, and if $\varphi(x)$ is strictly monotone in this interval, then $z = \varphi(x)$ can be introduced as a variable of integration on that interval, transforming S into

* Method of Stationary Phase Journal of Math. & Physics, vol 24, no 3-4 - 1945

$$\int_I f(x) e^{ip\varphi(x)} dx = \int_I f[\varphi^{-1}(z)] e^{ipz} dz$$

If, in addition to the above, $\varphi(x)$ and $\varphi''(x)$ are continuous and if $f(x)$ and $f'(x)$ exist and are continuous, this last integral can be integrated by parts, giving

$$S = \left\{ \frac{f[\varphi^{-1}(z)] e^{ipz}}{ip} \right\} - \frac{1}{ip} \int_I e^{ipz} \frac{d}{dz} f[\varphi^{-1}(z)] dz$$

and showing that on any such interval I ,

$$S \approx O\left(\frac{1}{p}\right).$$

Thus, condition (1) will be satisfied if, in the neighborhood of the stationary point α , the contribution to the integral is greater than $O\left(\frac{1}{p}\right)$.

This is clearly the case when the asymptotic formula (26) is valid, since there the dependences on p is as $1/\sqrt{p}$. Now it can be shown that (26) is valid whenever

$$\varphi(\alpha) = 0, \varphi''(\alpha) \neq 0 \text{ and } \varphi''(x) \text{ and } f[\varphi^{-1}(z)]$$

are of bounded variation in the neighborhood of the stationary value. Thus, to recapitulate, under these conditions, the maximum contribution comes from the stationary point and depends on p as $1/\sqrt{p}$, while the points which are not near the stationary point contribute terms depending upon p only as $1/p$.

To conclude this brief appendix, it should be remarked that Winter gives an extension of (10) which is valid under the same condition of $f[\varphi^{-1}(z)]$ if the first n derivatives of $\varphi(x)$ vanish at some point α while $\varphi^{n+1}(x)$ does not. These results could be used to extend the treatment of the high-pass case given above to the cases in which $\alpha^2 + 2b = 0$, etc.

C. L. DOLPH
C. E. SHANNON

Att.
B-392415 to 392428

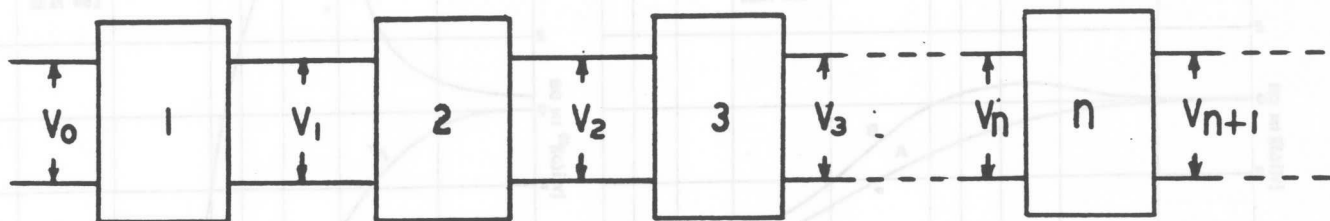


FIG. 1

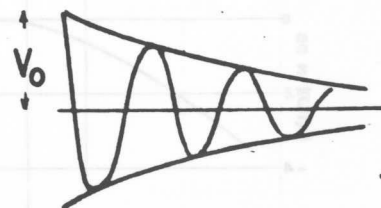


FIG. 2

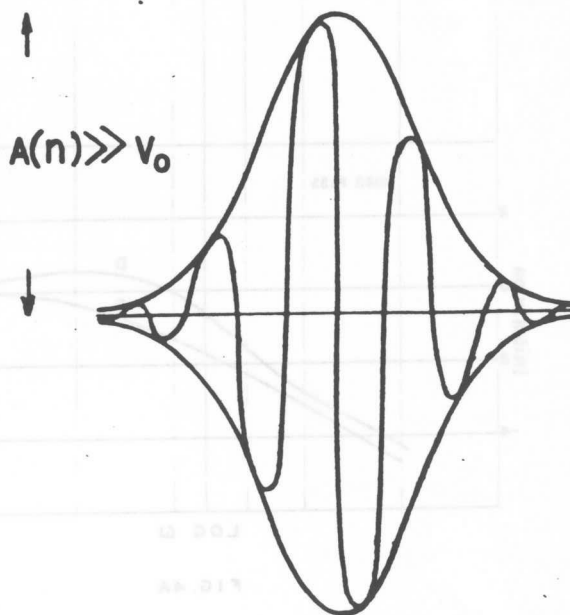


FIG. 3

BA-392415

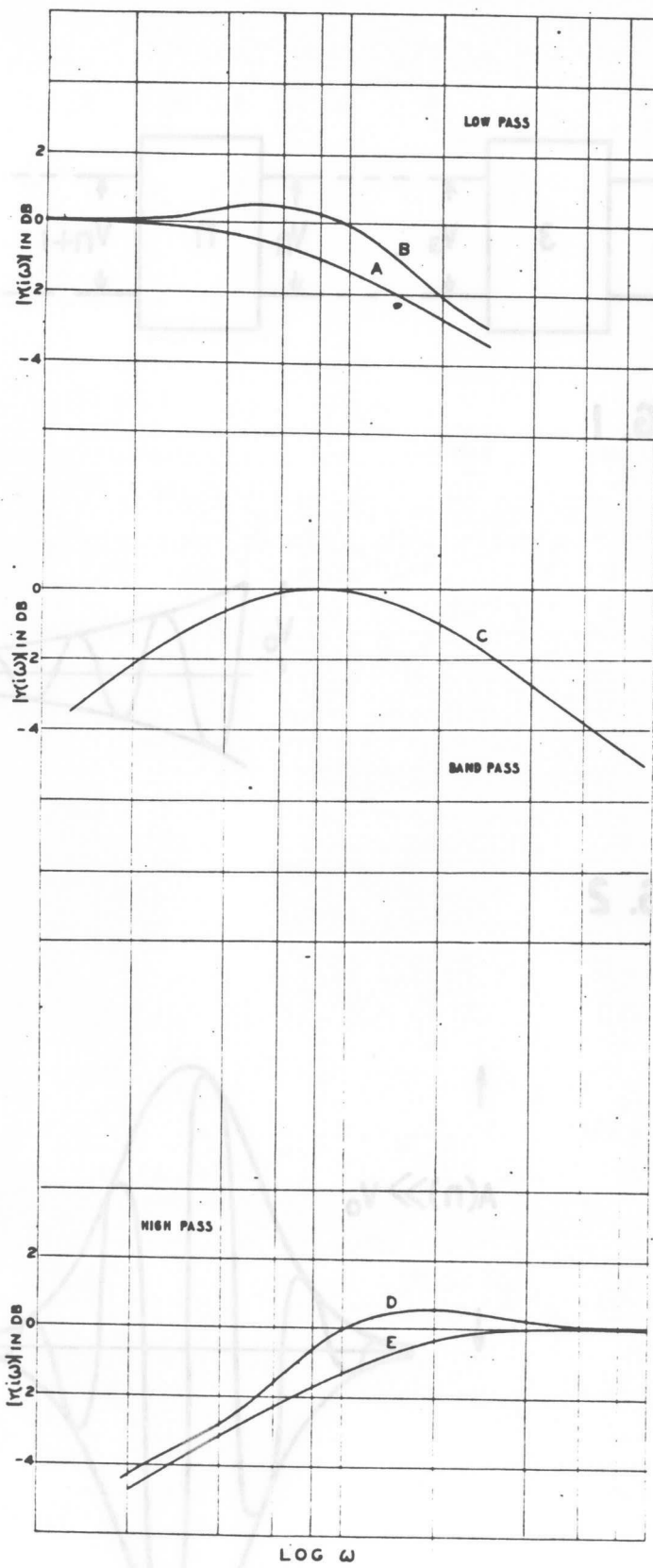


FIG. 4A

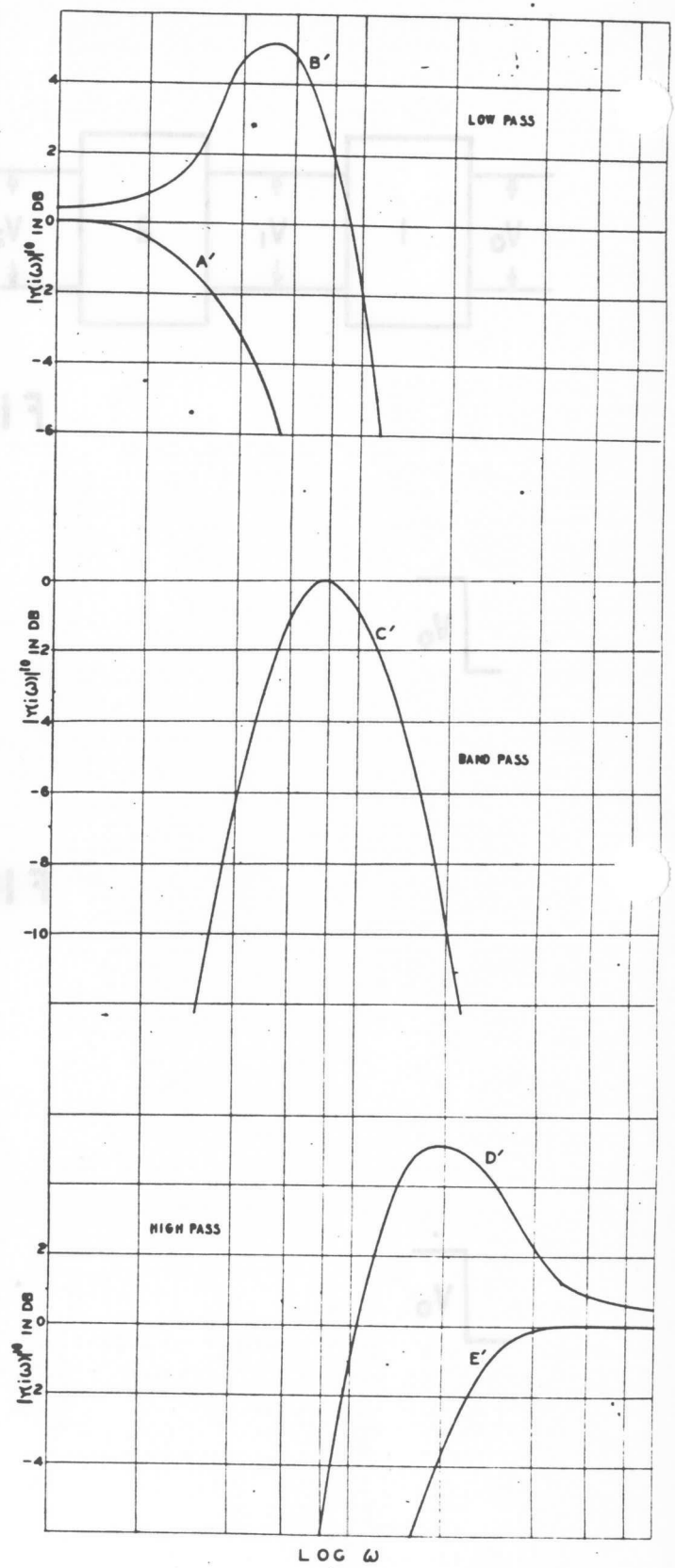


FIG. 4B

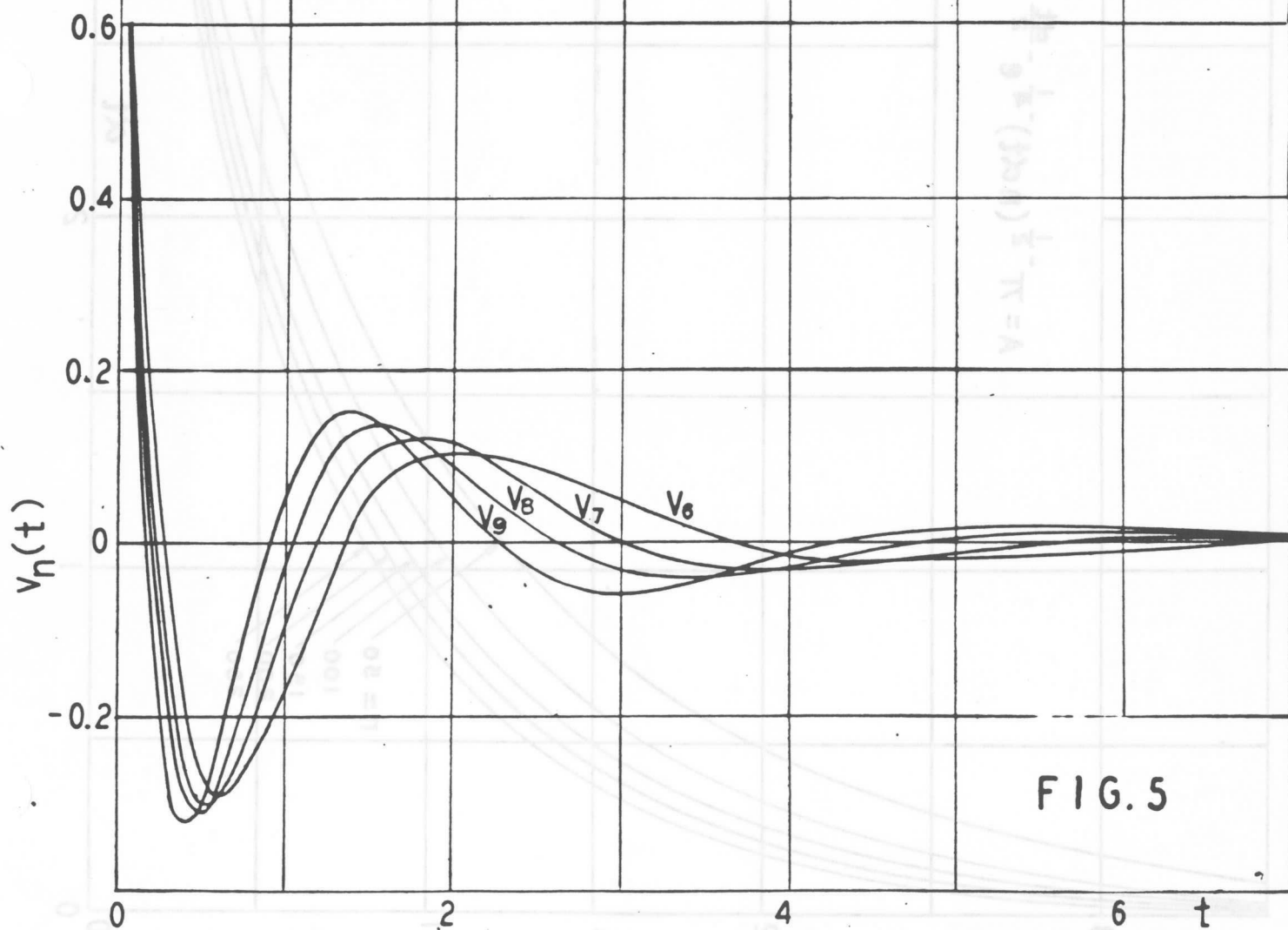
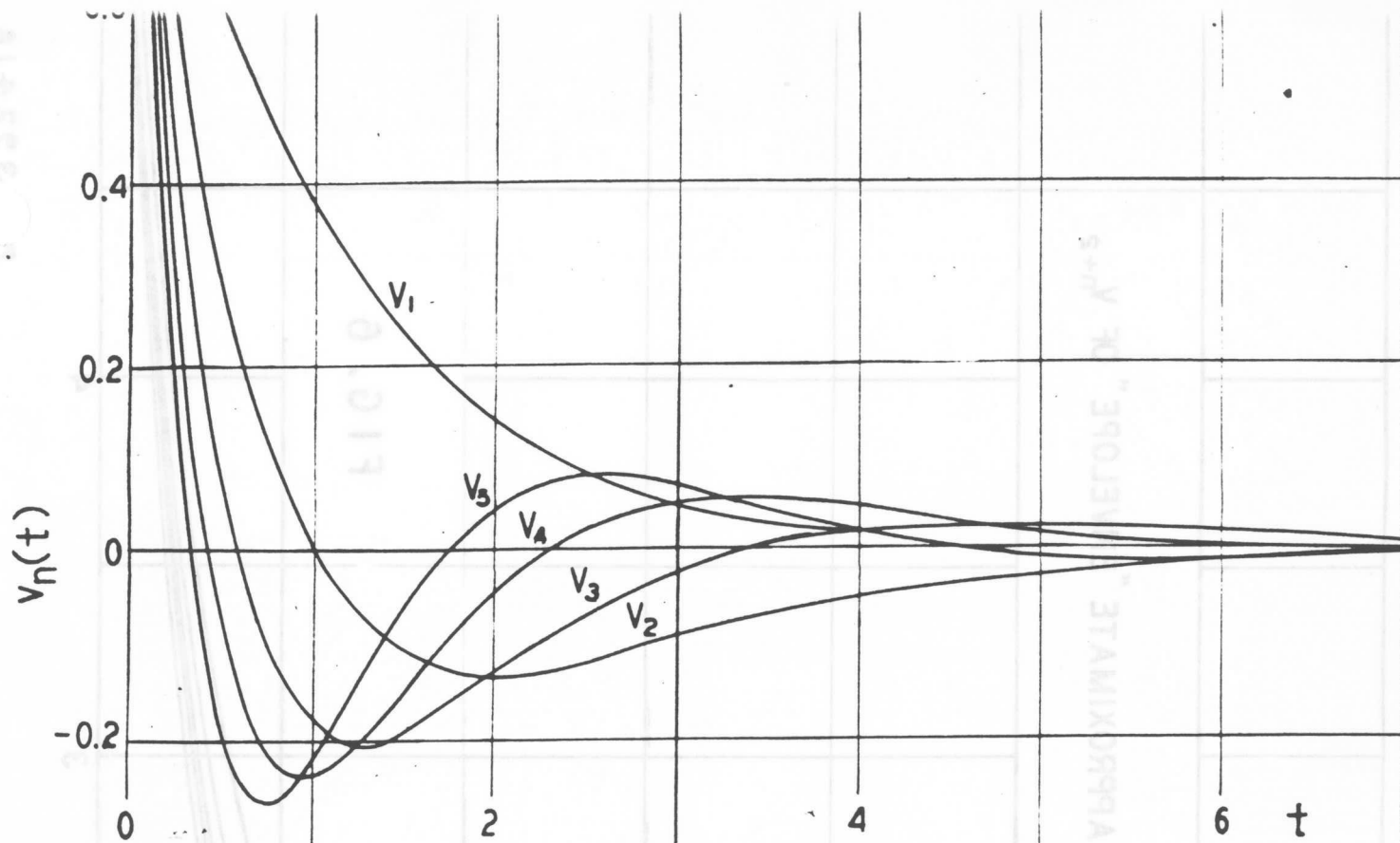


FIG. 5

$$A = \pi^{-\frac{1}{2}} (n\alpha t)^{-\frac{1}{4}} e^{-\frac{1}{2} - \frac{\alpha t}{2}} = \text{APPROXIMATE "ENVELOPE" OF } V_{n+2}$$

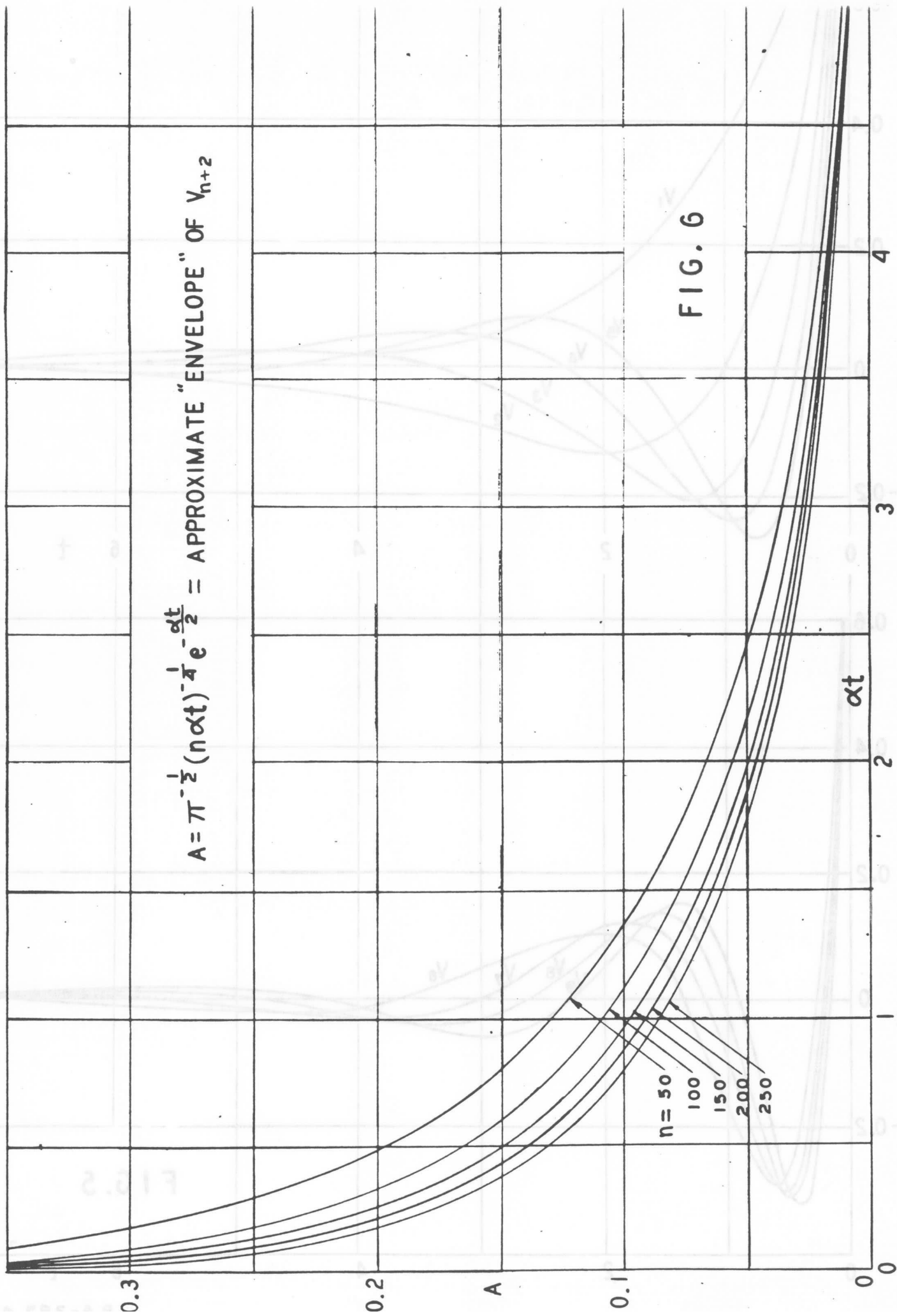


FIG. 6

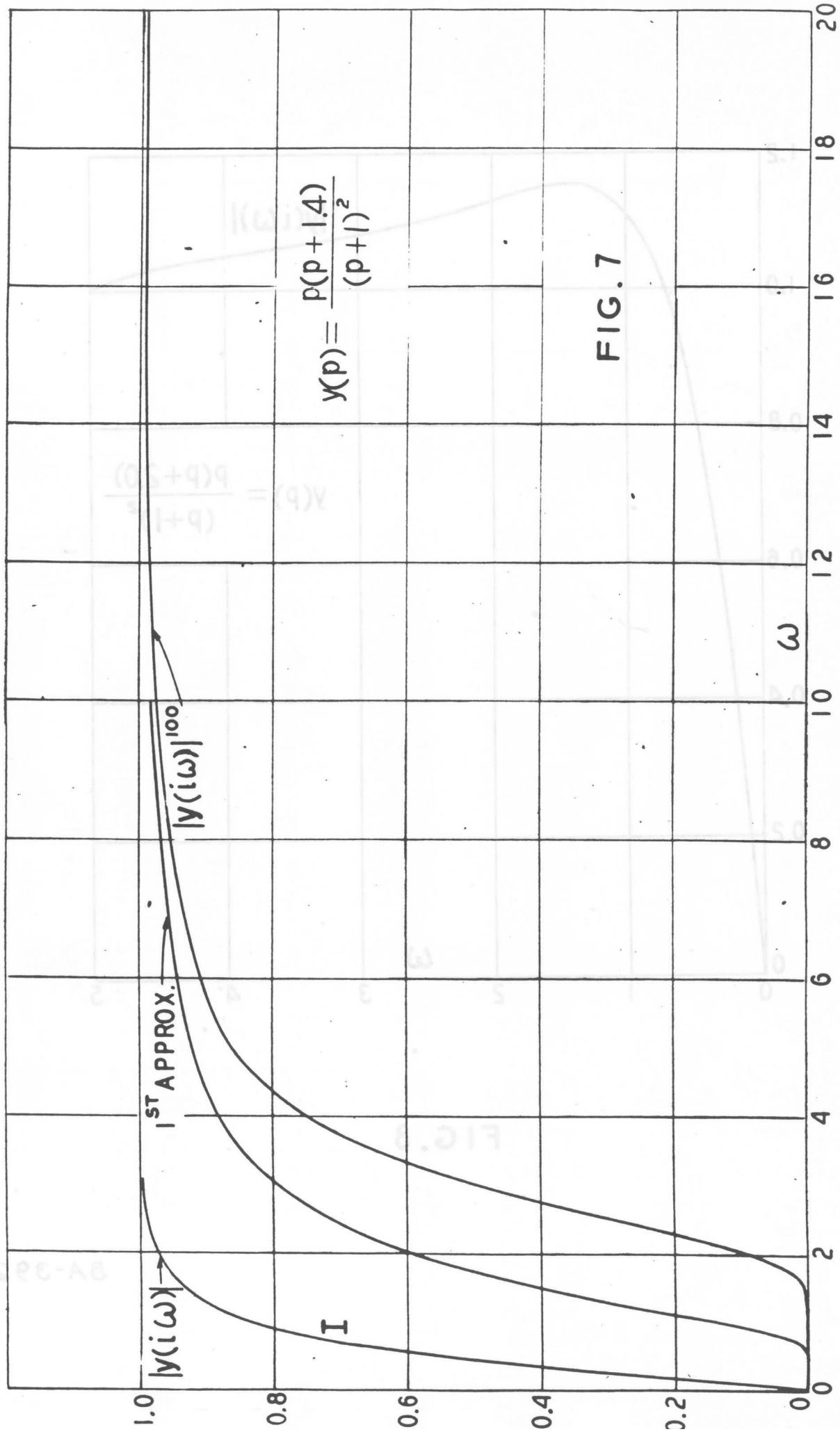


FIG. 7

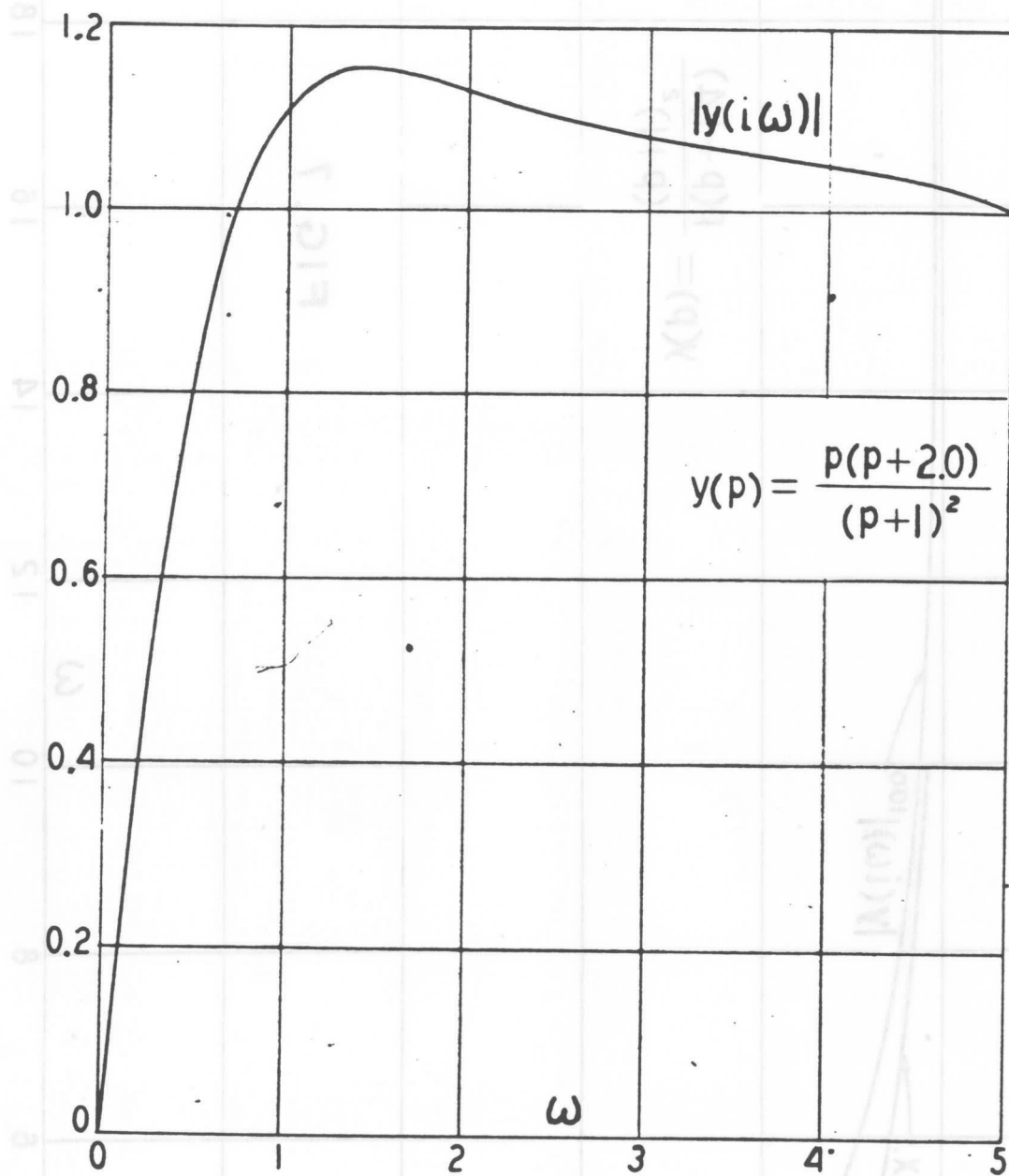
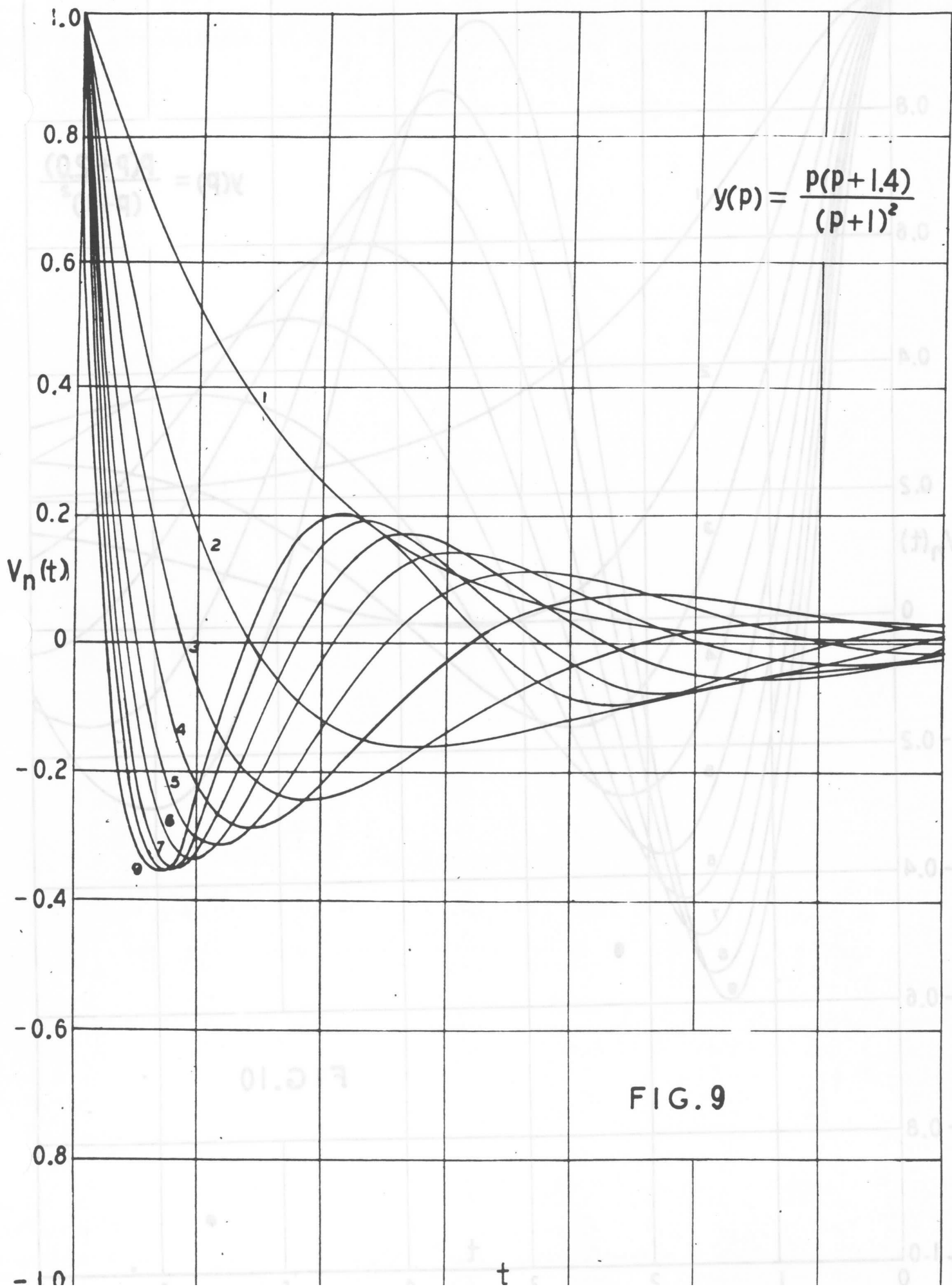
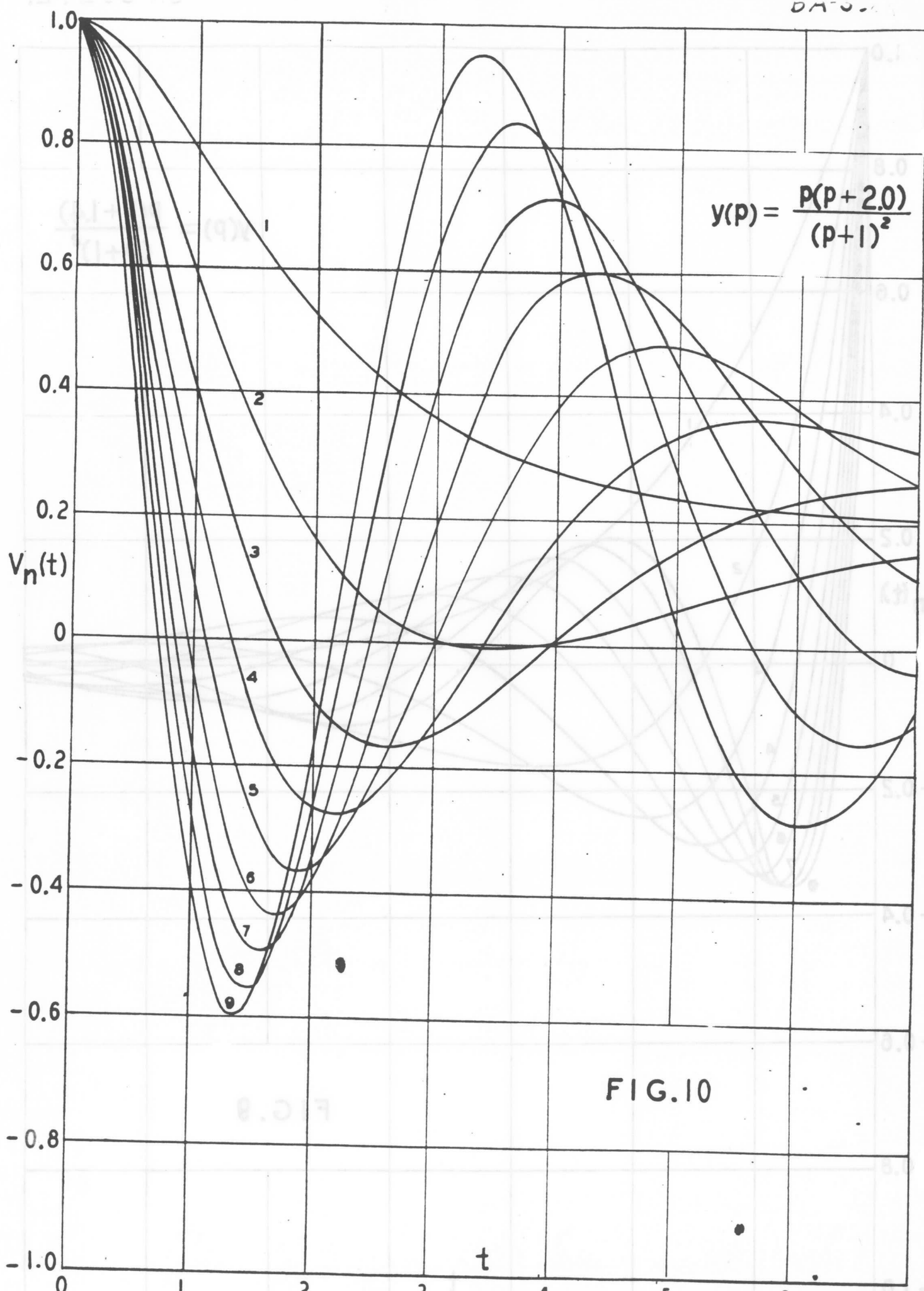


FIG. 8

BA-39242C





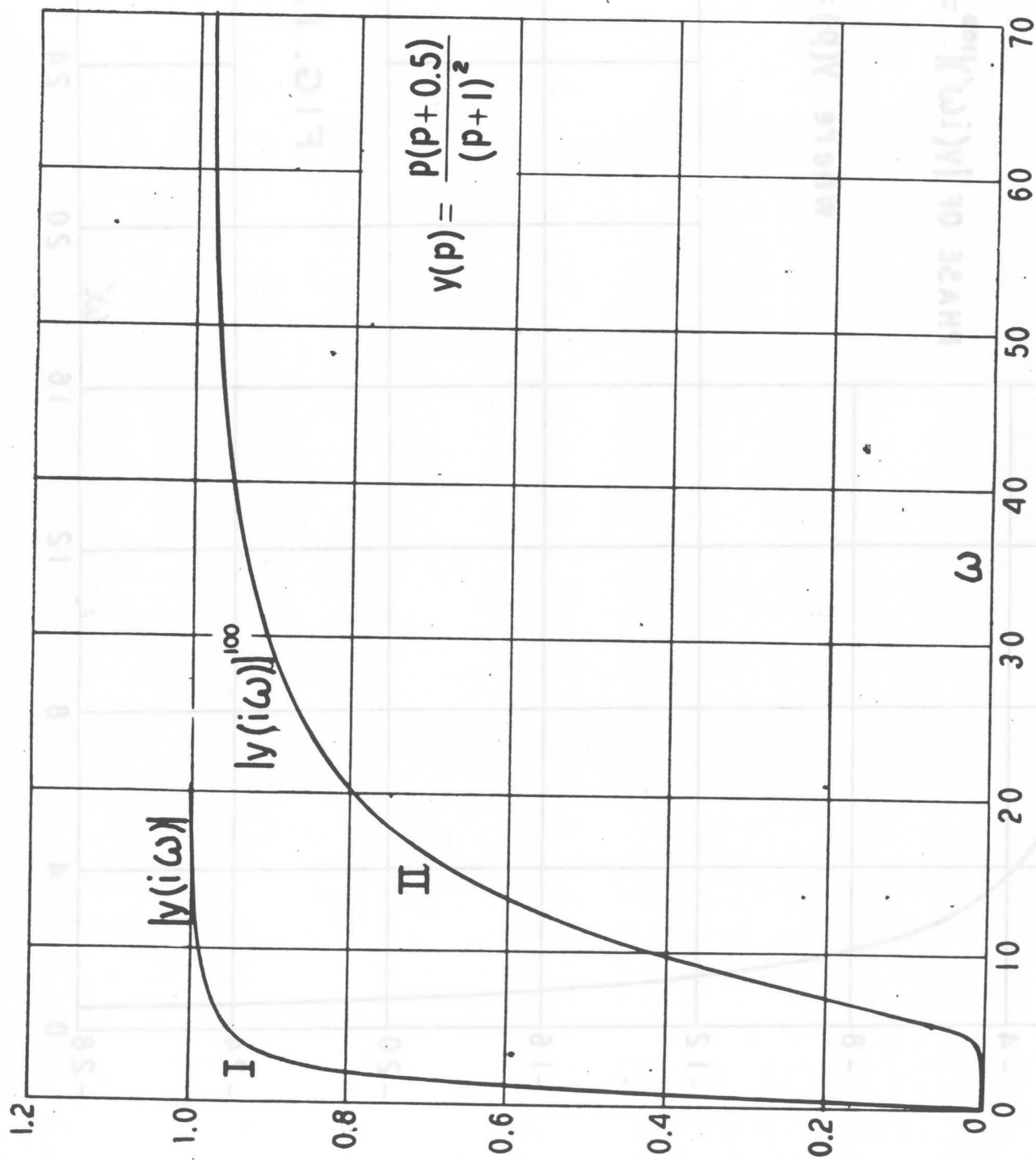
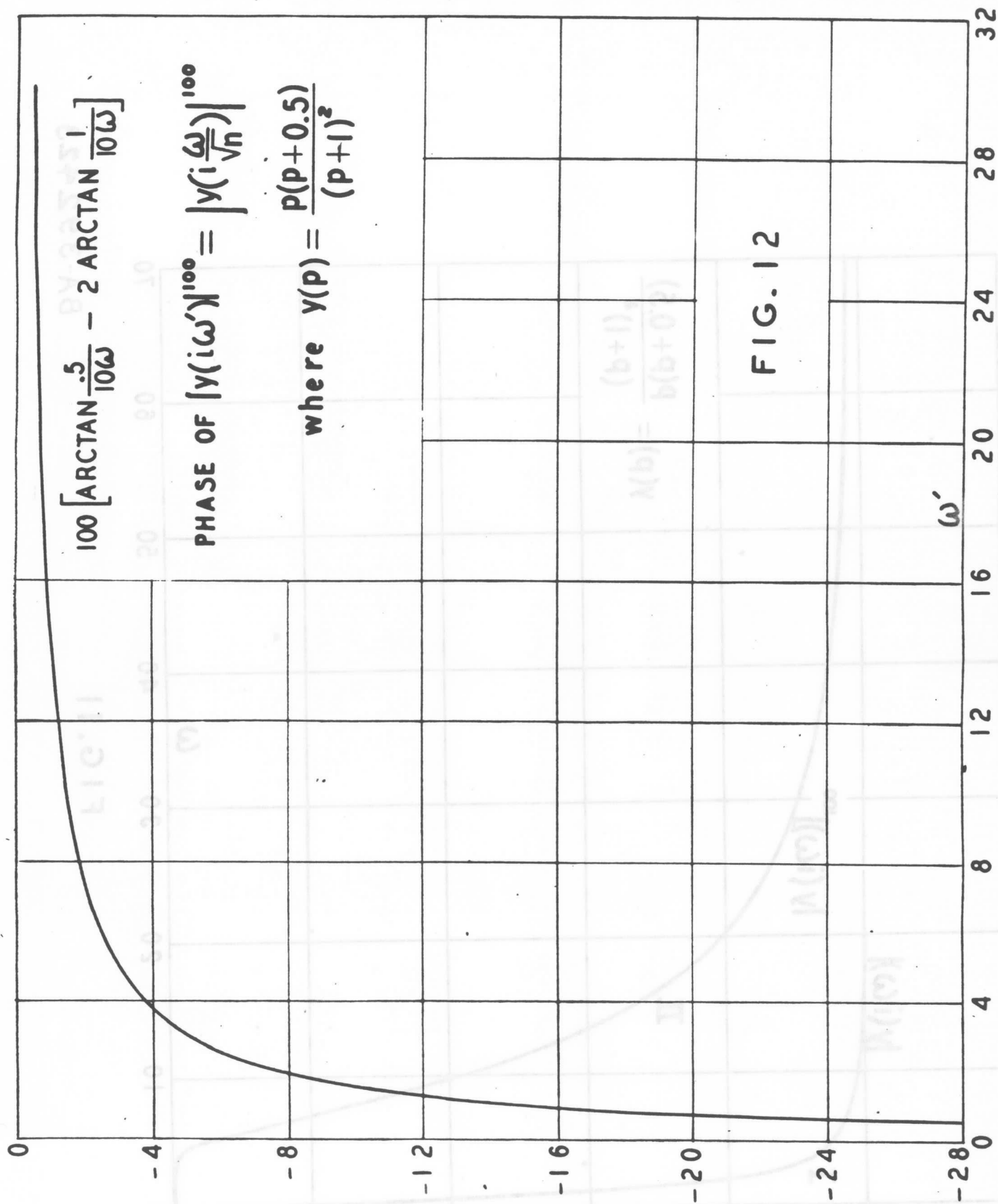
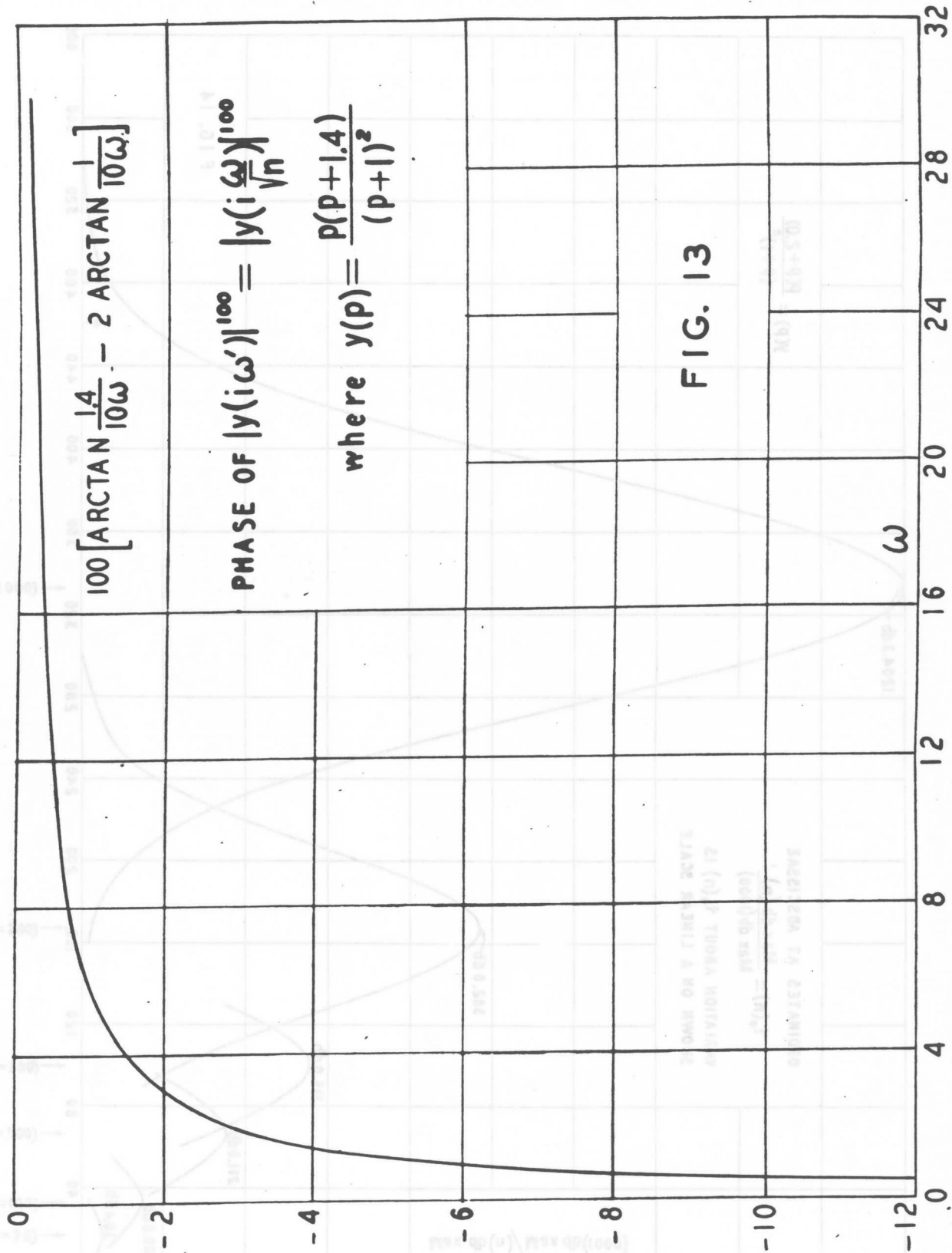
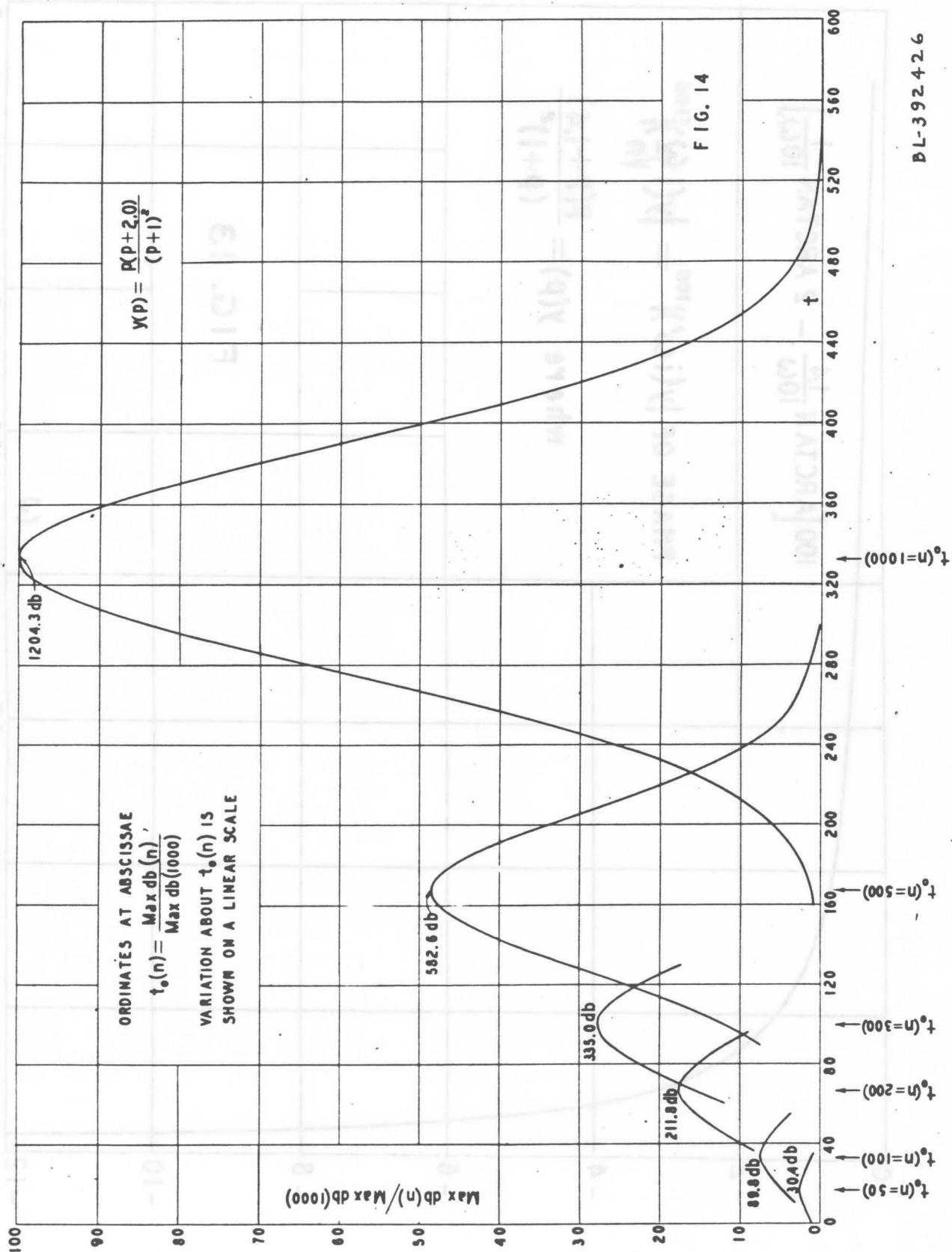


FIG. 11

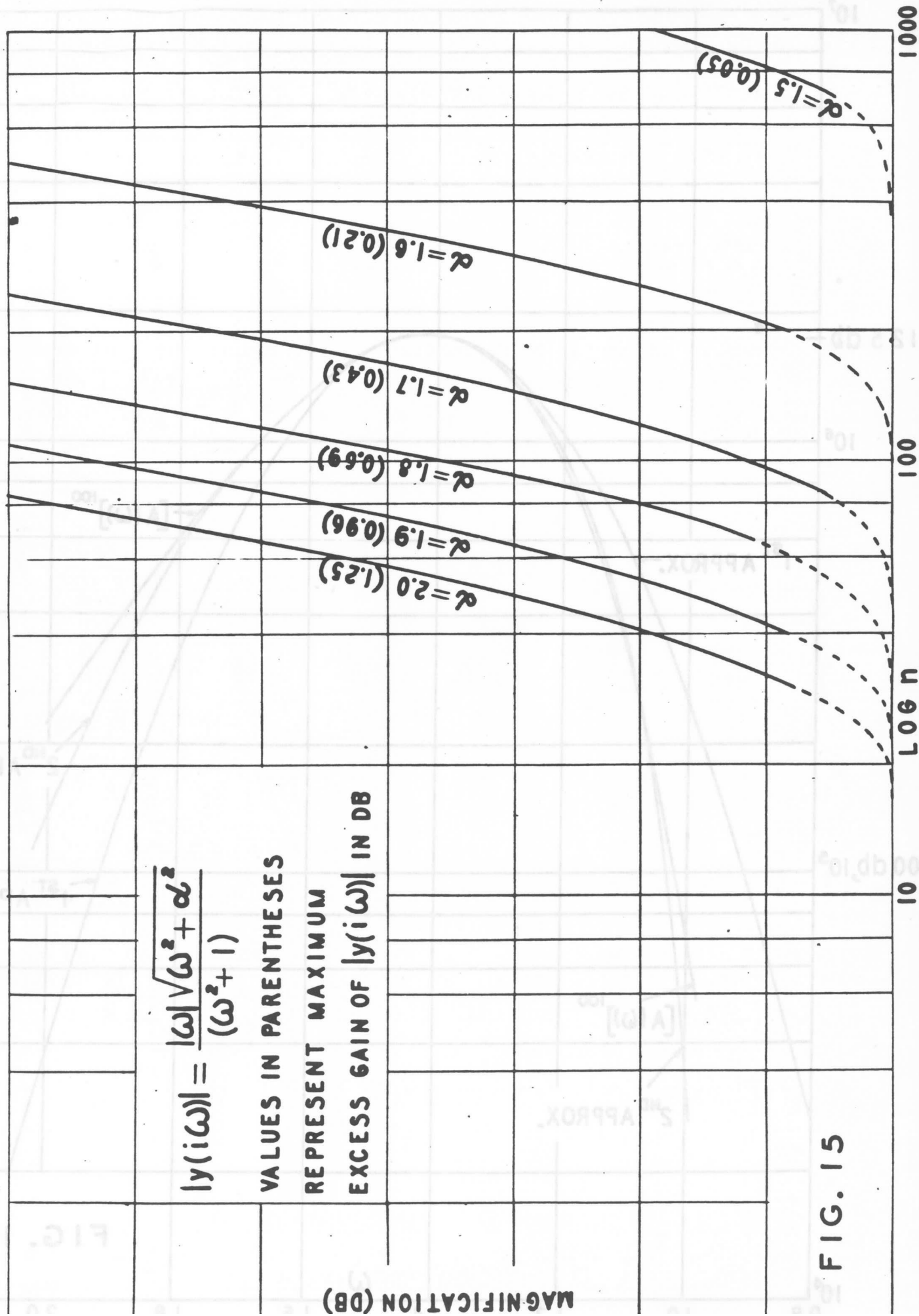
BA-392423







BL-392426



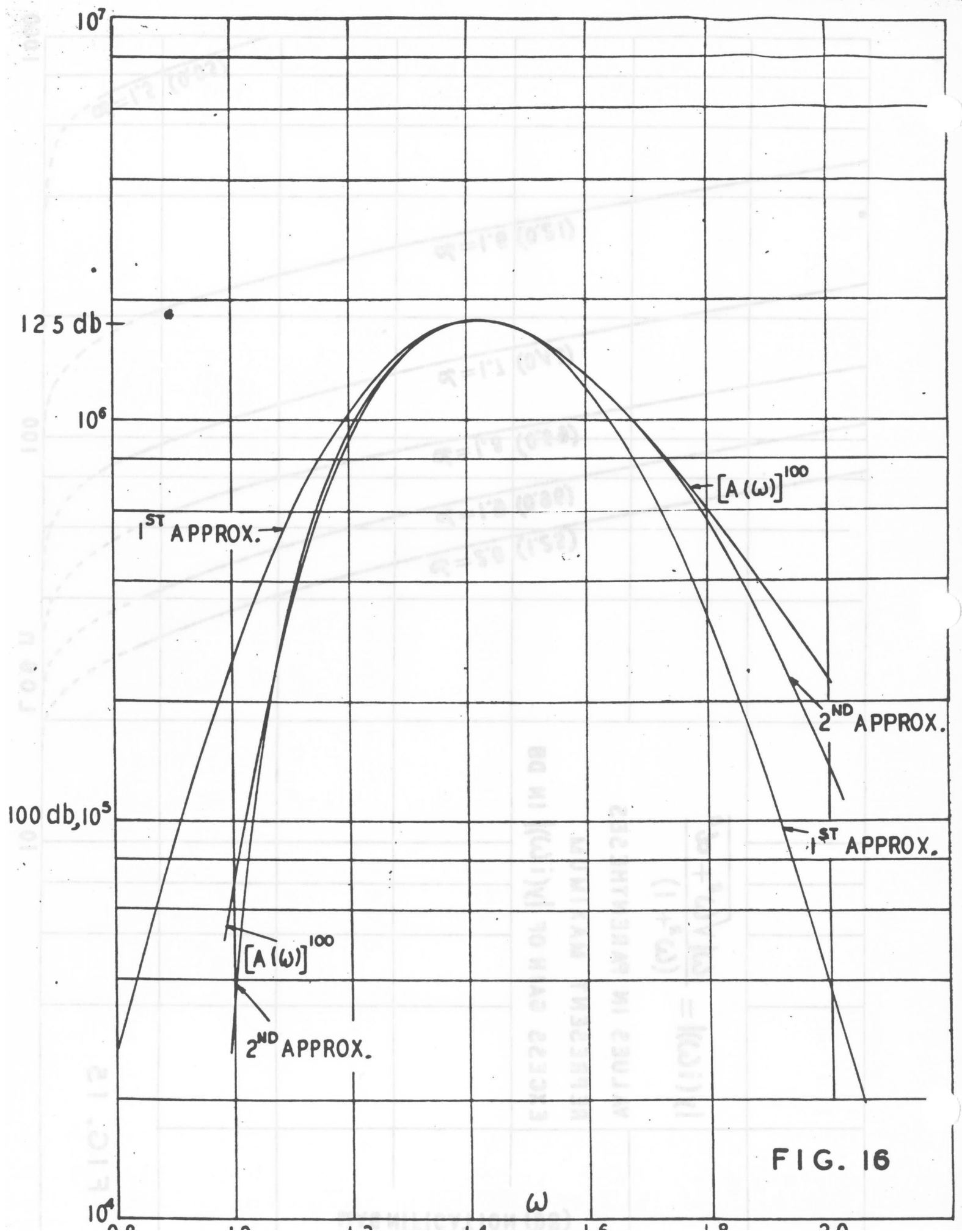


FIG. 16

Electronic Methods in Telephone Switching

C. E. Shannon

In the recent development of electronic digital computing machines various new tubes and other electronic devices have been designed which may be of use in machine switching. In particular the "selectron" tube developed by R. C. A. and the mercury acoustic delay tank provide large cheap memory devices in which information can be registered or read off in electronic time intervals (of the order of microseconds). Since one of the chief functions of the relays and switches in a telephone exchange is that of memory (e.g. the relays remember which calling and called lines should be connected together) it is worth while considering the possibility of using such tubes to replace ordinary electro-mechanical switching equipment.

Suppose we have an exchange (or set of exchanges) serving n subscribers and that the exchange can handle a peak load of m simultaneous conversations. These may be between any m pairs of the subscribers. Thus the exchange must be capable of assuming as many different states as there are of selecting m pairs of objects from n . This can be done in

$$\frac{n!}{m! 2^m (n - 2m)!}$$

different ways. For n and m large the logarithm of this is approximately $2m \log n$. If the logarithm is to the base ten then this is the required memory capacity of the exchange measured in decimal digits. If the logarithmic base is two the units are

binary digits. A single two-position relay has a capacity of $\log 2$ units (one binary digit or .30103 decimal digits), while S relays have $S \log 2$ units. A 10×10 crossbar switch has a capacity of $10 \log 10$, while a single commutator on a panel has capacity $\log r$, where r is the number of vertical positions of the brushes. Hence the number of relays required for a pure relay exchange would be

$$\frac{2m \log n}{\log 2},$$

the number of 10×10 crossbars would be

$$\frac{2m \log n}{10 \log 10},$$

etc. To these estimates must be added the losses due to inefficient use of the memory and also the memory of equipment used for functions other than merely remembering which connections are being held at a given time.

An ordinary relay is capable of remembering (by a holding circuit) one binary digit. A pair of vacuum tubes in a flip-flop circuit has the same memory capacity. The cost of these is of comparable magnitude, and thus if one designed an electronic telephone exchange by merely changing relays to equivalent vacuum tube circuits the chief advantage of the electronic circuit would be one of speed, an improvement of order 10^3 . In many cases this could produce a reduction of cost since frequently many identical units of a certain type must be supplied because the individual units are slow. This is apt to be the case with units which are associated with the beginning or end of calls but need not be used during the conversation. On the other hand equipment to be used throughout the call would offer less advantage under this tube for relay replacement since the expected duration of calls is long compared to electronic times.

The newer electronic memory devices, however, change this picture considerably. A selectron tube (when these tubes are in production) may be expected to cost \$100 or less depending on the demand. It is capable of holding 4096 binary digits, giving a cost per binary digit of the order of 2.5 cents, while the cost of the equivalent relay may be of the order of 2.5 dollars. Mercury delay lines can store information at a comparable cost. Thus it is not impossible that a reduction of the order 100 to 1 in switching equipment cost might be possible by the use of electronic devices, even in the parts where information must be stored for long periods of time.

An indication of how such tubes may be used is given in the attached figure. Fig. 1 is a block diagram of a simplified exchange. The calling parties are connected to an electronic commutator which samples the speech signals periodically and puts the various lines in the time division multiplex. The called parties are also connected in time division multiplex to a single channel by means of an electronic commutator or distributor. The function of the middle part is to rearrange the samples in such a way as to provide any desired interconnection between calling and called parties. This is done by dividing the sampling period into two equal parts. During the first half the signal plate of the upper selectron is connected by gate 1 into the calling line multiplex channel. Its windows are caused to open in sequence. Thus at the end of the first half-cycle the first samples of all the incoming channels have been written on the face of the tube in their regular order. During the second half-cycle gates 1 and 3 are closed and gates 2 and 4 are opened. Thus the output of the selectron is fed into the called line multiplex and the windows of the selectron are controlled by the other selectron tube 2. This tube has registered in a suitable notation the numbers of the

called line desired by the calling line. The windows of this tube are opened sequentially by the cycling unit and the numbers registered there control the windows on tube 1 allowing the sample from calling channel 1 to go into the proper place in the called line TDM.

By a more elaborate system it is possible to make use of the fact that only a small fraction of the lines will be busy at a given time, as is done in ordinary relay switching. This can be achieved by only supplying enough places in the distributors for the peak load. When a call originates the calling and called parties are assigned idle spaces in the distributor. The place assigned to the called party is registered in the selectron register corresponding to the place assigned to the calling party.

[32]

Some Generalizations of the Sampling Theorem

We have seen that a function of time $f(t)$ containing no frequencies over W cycles per second can be described by giving its value at Nyquist intervals (spaced $\frac{1}{2W}$ seconds apart). It can be reconstructed from these samples using the basic functions $\sin 2\pi Wt/2\pi Wt$, together with the same function shifted by integer numbers of Nyquist intervals. We now consider some generalizations of this result.

In the first place the particular function $\sin 2\pi Wt/2\pi Wt$ is by no means necessary for the reconstruction. In fact any function $\varphi(t)$ which contains all frequencies up to W is satisfactory. More precisely the spectrum of $\varphi(t)$ should not vanish over any finite set of frequencies (set of positive measure) up to W . If $\varphi(t)$ satisfies this condition the original function $f(t)$ can be reconstructed using $\varphi(t)$ and its shifted images $\varphi(t + \frac{K}{2W})$. That is coefficients a_K can be found such that

$$f(t) = \sum_{K=-\infty}^{\infty} a_K \varphi(t + \frac{K}{2W}) .$$

In general the coefficients are not found as easily as in the special case where $\varphi(t) = \sin 2\pi Wt/2\pi Wt$ (when they are merely the values of $f(t)$ at the Nyquist points) but they may be calculated as follows. Let $F(\omega)$ be the spectrum of $f(t)$ and $\Phi(\omega)$ be the spectrum of $\varphi(t)$. Expand the function $F(\omega)/\Phi(\omega)$ in a Fourier series using $-W$ to $+W$ as the fundamental interval.

Thus

$$\frac{F(\omega)}{\Phi(\omega)} = \sum a_K e^{i \frac{K\omega}{2W}}$$

or

$$F(\omega) = \sum a_K \Phi(\omega) e^{i \frac{K\omega}{2W}}.$$

Taking the transform of the equation we obtain the desired expansion

$$f(t) = \sum a_K \varphi(t + \frac{K}{2W}).$$

The coefficients in the expansion can therefore be determined as the coefficient of a Fourier series expansion of $F(\omega)/\Phi(\omega)$. In general the function $\varphi(t + \frac{K}{2W})$ will not form an orthogonal set and therefore the energy in $f(t)$ cannot be found from $\sum a_K^2$ as it was in the simple case where $\varphi(t) = \sin 2\pi Wt/2\pi Wt$.

A physical method of performing this expansion can also be given. Consider a filter which gives the output $\sin 2\pi Wt/2\pi Wt$ when the input is $\varphi(t)$. If the function $f(t)$ is passed through this filter the amplitudes of the output at Nyquist intervals will be the desired coefficients. This is true since this output can be considered as expanded in the functions $\sin 2\pi Wt/2\pi Wt$ with the amplitudes as coefficients, and the inverse filter would restore the original function and change each of these functions with $\varphi(t)$ at the corresponding Nyquist point.

A function $f(t)$ can also be determined from a knowledge of its value and derivative at alternate Nyquist points:

$$f\left(\frac{2K}{2W}\right) \quad \text{and} \quad f'\left(\frac{2K}{2W}\right) .$$

We have here the same number of measurements per second, $2W$, but half of these are ordinates of $f(t)$ and half are derivatives. The reconstruction of $f(t)$ from these values can be carried out simply using two basic functions:

$$\varphi_1(t) = \frac{\sin^2 \pi W t}{(\pi W t)^2}$$

$$\varphi_2(t) = \frac{\sin^2 \pi W t}{(\pi W t)} .$$

Both of these lie entirely within the band W and φ_1 has the property that it and its first derivative vanish at alternate Nyquist points (except for $t = 0$ where the function is 1 and its first derivative 0). Likewise φ_2 and φ_2' vanish at alternate Nyquist points except at $t = 0$ where $\varphi_2 = 0$ and $\varphi_2' = 1$. Thus we can fit the ordinates of the original function $f(t)$ using φ_1 and its shifted images (shifted by two Nyquist intervals). The derivatives of $f(t)$ are fitted using φ_2 and its shifted images. Due to the vanishing of these functions none of the fittings interfere. The function constructed by this process must lie within the band and have the same values and derivatives as the original function $f(t)$ at alternate Nyquist points. That there is only one such function can be shown by arguments similar to those used in the basic sampling theorem, generalized by breaking down the spectrum into an even and an odd part.

It is possible to carry this further and determine a function from knowledge of its value and first $(n - 1)$ derivative at points separated n Nyquist intervals apart. In this case the basic functions are

$$\varphi_1 = \frac{\sin^n \left(\frac{2\pi Wt}{n} \right)}{\left(\frac{2\pi Wt}{n} \right)^n}$$

$$\varphi_2 = \frac{\sin^n \left(\frac{2\pi Wt}{n} \right)}{\left(\frac{2\pi Wt}{n} \right)^{n-1}}$$

$$\varphi_3 = \frac{\sin^n \left(\frac{2\pi Wt}{n} \right)}{\left(\frac{2\pi Wt}{n} \right)^{n-2}}$$

...

$$\varphi_n = \frac{\sin^n \left(\frac{2\pi Wt}{n} \right)}{\frac{2\pi Wt}{n}}$$

These functions possess the properties:

1. They lie within the band W .
2. They vanish at $t = \frac{Kn}{2W}$ $K = \pm 1, \pm 2, \dots$,
(that is at n -th Nyquist points) and also their
1st, 2nd, ..., $(n-1)$ derivatives.
3. At $t = 0$, all derivatives of φ_s vanish except the s -th
derivative which is 1.

Consequently we can reconstruct $f(t)$ by using φ_s to adjust the s derivatives ($s = 0, 1, \dots, n-1$) and these adjustments will not interfere.

The functions φ_s and their spectra are shown in Fig. 1 for the cases $n = 1, 2, 3$.

C. E. SHANNON

Att.
Figure 1

March 4, 1948

[34]

The Normal Ergodic Ensembles of Functions

Among the possible probability distributions in a one-dimensional space certain ones are of special importance because of their simple mathematical properties and frequent occurrence in the physical world. The most important of these is the normal or Gaussian distribution with a density function:

$$1/\sqrt{2\pi} \sigma \exp \left[-\frac{1}{2} x^2/\sigma^2 \right]$$

In an n-dimensional space the most important distribution function is an n-dimensional generalization of this, the n-dimensional normal distribution:

$$|a_{ij}|^{\frac{1}{2}}/(2\pi)^{\frac{n}{2}} \exp \left[-\frac{1}{2} \sum a_{ij} x_i x_j \right]$$

Here a_{ij} is the associated quadratic form and $|a_{ij}|$ the determinant of this form. This form is positive definite and the surfaces of the constant probability are found by setting the argument of the exponential function equal to a constant

$$\sum a_{ij} x_i x_j = C$$

and are therefore coaxial ellipsoids in the space. The directions of the axes of this ellipsoid are those of the eigenvectors of the form a_{ij} and the lengths are inversely proportional to the corresponding eigenvalues. By a rotation of axes the new coordinate system can be lined up with these directions and the distribution function reduced to

$$(\lambda_1, \lambda_2, \dots, \lambda_n)^{\frac{1}{2}} (2\pi)^{-\frac{n}{2}} \exp - \frac{1}{2} \sum \lambda_i y_i^2$$

where the λ_i are the (positive) eigenvalues and the y_i are the new coordinates. The form a_{ij} being positive definite has an inverse A_{ij} which is also positive definite with eigenvalues $\mu_i = \lambda_i^{-1}$.

The properties of the n-dimension normal distribution which give it particular mathematical importance are the following.

1. If x_i and y_i are two chance vector variables, which are independent and distributed according to n-dimensional normal distributions with quadratic forms a_{ij} and b_{ij} (inverses A_{ij} and B_{ij}), then the chance vector variable $z_i = x_i + y_i$ is also distributed normally with the form c_{ij} , whose inverse is $C_{ij} = A_{ij} + B_{ij}$.

2. If x_i is a normally distributed vector variable and $y_j = \sum_i r_{ij} x_i$ is a vector variable which is a linear operation on x_i (possibly of smaller dimension than n) then y_j is normally distributed with the inverse form

$$B_{ij} = \sum_{s,t} r_{is} r_{jt} A^{st}.$$

3. Under certain quite broad conditions the resultant of a large number of small chance vector variables, x_i^s ($s = 1, 2, \dots, N$) with arbitrary distribution functions, which are independent gives a normal distribution for

$$y_i = \sum_s x_i^s$$

with

$$B_{ij} = \sum_s A_{ij}^s, \quad A_{ij}^s = \overline{x_i^s x_j^s}$$

providing no term of the sum contributes more than a small fraction to any B .

4. If the a priori probabilities for each of two independent vectors x_i and y_i are both normal, the a posteriori probability of x_i when we know the sum $x_i + y_i = r_i$ is normally distributed (about a displaced mean, however).

5. The mean value of $x_i x_j$ for x_i normal is given by

$$\overline{x_i x_j} = A_{ij}.$$

Among the many possible ergodic ensembles of functions $f_a(t)$ there is also a certain class of particular mathematical and physical importance. This class of ensembles can be considered a generalization of the n -dimensional normal distribution to infinite dimensional function spaces ergodic under translations in time. We shall call these normal ergodic ensembles of functions. They are completely specified by giving their power spectra $P(\omega)$ or their autocorrelation functions $A(t)$ which are the Fourier transforms of the power spectra. The normal ergodic ensembles can be defined in various ways. They occur physically when we pass a thermal noise through a filter, shaping the power spectrum to $P(\omega) = |Y(\omega)|^2$, $Y(\omega)$ being the admittance of the filter.

In the literature on noise these ensembles are often treated in a loose somewhat illogical fashion by using either of two "representations." The first representation is

$$\sum_{n=0}^{\infty} |P(n\Delta f)\Delta f \cos (n\Delta f t + \theta_n) |.$$

The θ_n are all uniformly and independently distributed over all values from 0 to 2π . This representation amounts to making the noise the sum of a large number of small sinusoidal waves with random phases, and amplitudes adjusted to give the proper power density in any small frequency range. The frequency increment between adjacent waves Δf is supposedly very small and in use one evaluates any desired statistic of this set of functions and determines the limit approached by this statistic as $\Delta f \rightarrow 0$. This limit is taken to be the desired statistic of the normal ergodic ensemble. The second representation is similar but uses normally distributed amplitudes a_n whose variance σ^2 is equal to $P(\omega)$

$$\sum a_n \Delta f \cos (n\Delta f t + \theta_n) .$$

Actually these "representations" will not give the correct answer in all cases. For example, if we ask what fraction of the functions in the representation ensemble $r_{\Delta f}$ are periodic, we find that all are, so the probability is unity, and the limit as $\Delta f \rightarrow 0$ is also therefore unity, while almost none of the functions in the ergodic normal ensemble are periodic. However it can be shown that if we restrict ourselves to what we

have called physical statistics, the answer will be identical; the normal ergodic ensemble is the physical limit of either of the above ensembles as $\Delta f \rightarrow 0$.

A more logical definition of a normal ergodic ensemble can be given as follows. We divide the frequency range up into unit intervals and construct the sequence of "flat" ensembles for these intervals. These will be given by

$$\sum a_n \sin nt .$$

These ensembles are passed through shaping filters to give the proper power spectrum in the interval in question and the results added.

The normal ergodic ensembles have properties analogous to the n -dimensional normal distributions which we have given. We have

Theorem: The sum of two functions $f_\alpha(t) + g_\beta(t)$ where f and g are from normal ergodic ensembles with spectra P_1 and P_2 is normal ergodic with spectrum $P_1 + P_2$.

Theorem: The output of any linear invariant transducer driven by a normal ergodic ensemble is normal ergodic with spectrum $|Y(\omega)|^2 P(\omega)$.

Theorem: Any finite dimensional linear operation on a normal ergodic ensemble gives a normally distributed vector.

C. E. SHANNON

March 15, 1948

[35]

Systems Which Approach the Ideal as $\frac{P}{N} \rightarrow \infty$

We will show that it is possible to construct an instantaneous system for sufficiently large $\frac{P}{N}$ for transmitting a sequence of binary digits such that the frequency of errors is arbitrarily small and the power required only slightly greater in db than the ideal for the corrected rate of transmission. More precisely we have the

Theorem: Given any $\epsilon > 0$ and $\delta > 0$ we can transmit binary digits on an instantaneous basis with frequency of errors $< \epsilon$ and corrected rate of transmission

$$R > W \log \left[1 + (1 - \delta) \frac{P}{N} \right]$$

The system to be used is of PCM type with an extremely large number of amplitude levels. Let there be 2^S levels, and number them with a binary notation, but in the Stibitz type code, so that only one binary digit changes on going to an adjacent level. If we are in error by d levels, at most d binary digits of the s will be incorrect. If there are many levels in the σ distance (\sqrt{N}) of the noise the expected number of errors will be approximately

$$a = \frac{1}{\sqrt{2\pi} \sigma} \int_{-\infty}^{\infty} \left| \frac{x}{d} \right| e^{-\frac{x^2}{2\sigma^2}} dx$$

We take $\frac{P}{N}$ large enough so that $\epsilon s > a$. Thus the frequency of errors in our final result will be $< \epsilon$. The levels should not be spaced uniformly but according to the density of a normal distribution. If this is done the received signal will be nearly Gaussian with $\sigma = \sqrt{P + N}$ and the corrected rate of transmission

$$R > W \log \left[1 + (1 - \delta) \frac{P}{N} \right].$$

C. E. SHANNON

March 29, 1948

Theorems on Statistical Sequences

If it is possible to go from any state with $P > 0$ to any other along a path of probability $p > 0$, the system is ergodic and the strong law of large numbers can be applied. Thus the number of times a given path p_{ij} in the network is traversed in a long sequence of length N is about proportional to the probability of being at i and then crossing this path, $P_i p_{ij} N$. If N is large enough the probability of percentage error $\pm \delta$ in this is less than ϵ so that for all but a set of small probability the actual numbers lie within the limits

$$(P_i p_{ij} \pm \delta) N$$

Hence the probability that nearly all sequences lie within limits $\pm \delta$ is given by

$$p = \pi p_{ij}^{(P_i p_{ij} \pm \delta) N}$$

and $\frac{\log p}{N}$ is limited by

$$\frac{\log p}{N} = \sum (P_i p_{ij} \pm \delta) \log p_{ij}$$

or

$$\left| \frac{\log p}{N} - \sum P_i p_{ij} \log p_{ij} \right| < \epsilon$$

Thus we have:

Theorem: For almost all sequences

$$\lim_{L \rightarrow \infty} \frac{-\log P}{L} = H = - \sum P_{ij} \log P_{ij}$$

where p is the probability of the sequence having the block of length L starting at the first position.

Thus for all but a set of blocks of probability $< \epsilon$ and for N large enough

$$(H - \eta)N < -\log p < (H + \eta)N$$

$$-p(H - \eta)N < -\sum p \log p < -p(H + \eta)N$$

where we have summed over all but the set of small probability ϵ .

$$\sum p(H + \eta)N \leq (H + \eta)N \sum p \leq (H + \eta)N$$

$$\text{and } \sum p(H - \eta)N \geq (H - \eta)N \sum p \geq (H - \eta)N (1 - \epsilon)$$

For the set of small probability

$$\left| -\sum p \log p \right| \leq \left| \sum \frac{\epsilon}{2^{RN}} \log \frac{2^{RN}}{\epsilon} \right|$$

since this is maximized for $\sum p = \epsilon$ by making all p equal, and the number of them $\leq \frac{1}{2^{RN}}$. But this is dominated by

$$\left| -\sum p \log p \right| \leq \left| \epsilon RN \log \frac{\epsilon}{2} \right|$$

$$\leq 0$$

with θ as small as desired for sufficiently large N and small ϵ . Hence this does not affect the sum in the limit as $N \rightarrow \infty$ and we have the

Theorem:

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum p(B_1) \log p(B_1) = H$$

where $p(B_1)$ is the probability of block B_1 of length L , and the sum is over all possible blocks.

We now prove the

$$\begin{aligned} \text{Theorem } H &= - \sum p(B_1 S_j) \log p_{B_1}(S_j) \\ &= \lim_{L(B) \rightarrow \infty} - \sum q(B_1 S_j) \log q_B(S_j) \end{aligned}$$

where $p(B_1, S_j)$ is the probability of block B_1 followed by S_j and $p_{B_1}(S_j)$ is the conditional probability of S_j after the block B_1 is known to occur. $q(B_1, S_j)$ is the probability when B_j is computed on the basis of any initial state probabilities, not necessarily the proper ones and $q_{B_1}(S_j)$ the corresponding conditional probabilities.

The first equality is true since we may sum first on all B_1 leading to a given state K . The terms $q_{B_1}(S_j)$ are then all equal to p_{Kj} and the terms $q(B_1 S_j)$ sum to $P_K p_{Kj}$ gives the desired result.

If the q 's are used, the $q_{B_1}(S_1)$ are still p_{kj} where k is the state in which B_1 ends.

$$q_{B_1 \rightarrow k}(B_1, S_j) = p_{kj} \frac{P(B_1)}{P(B_1 \rightarrow k)}$$

$$\rightarrow p_{kj} p_k$$

since any initial distribution tends toward equilibrium.

We have shown that apart from a set of small probability, the probabilities of blocks of length L lie within the limits

$$2^{-(H + \delta)N} < p < 2^{-(H - \delta)N}$$

where δ can be made small by taking N large enough. Let the maximum number of blocks of length N when we delete a set of measure ϵ be $G_\epsilon(N)$. Then:

$$\sum_{\text{remaining set}} p = (1 - \epsilon)$$

$$G_\epsilon(N) p_{\max} = G_\epsilon(N) 2^{-(H - \delta)N}$$

$$\log G_\epsilon(N) > (H + \delta)N + \log(1 - \epsilon)$$

Hence $\lim_{N \rightarrow \infty} \frac{\log G_\epsilon(N)}{N} = \varphi(\epsilon) \geq \delta$

Similarly

$$1 > \sum p > G_\epsilon(N) P_{\min} \\ = G_\epsilon(N) 2^{-(H - \delta)N}$$

from which we obtain

$$\frac{\log G_\epsilon(N)}{N} > H - \delta$$

and $\varphi(\epsilon) \geq H$

Hence we have

Theorem: $\varphi(\epsilon) = H$ for $\epsilon \neq 0, 1$

The fact that for large N nearly all blocks have a probability limited by

$$\left| \frac{-\log p}{N} + \delta \right| < \epsilon$$

does not imply that these probabilities approach equality.

In fact they will generally diverge from one another but the db range becomes small compared to N , since for p 's satisfying

this inequality

$$\frac{\log P_{\max}}{N} - \frac{\log P_{\min}}{N} = \frac{\log \frac{P_{\max}}{P_{\min}}}{N} < 2\epsilon$$

It is possible to show, however, that there exists among the blocks of length N a subset, all of equal probability which have the same growth with N as the set including all blocks except those of small probability totaling less than ϵ : namely, the subset will contain more than $2^{(H - \delta)N}$ elements with δ arbitrarily small.

Consider all blocks beginning in a given state, say state 1, and ending in this state. Let these blocks B_1, \dots, B_N, \dots have lengths $n_1, n_2, \dots, n_N, \dots$ and conditional probabilities $p_1, p_2, \dots, p_N, \dots$, when we start from state 1. We first prove

Theorem: $\sum p_i n_i = P_1^{-1}$

$$P_1 \sum_1^{\infty} p_i \log p_i = -2$$

The first part is true since the ergodic character of the system makes the inverse frequency of occurrence of state 1, P_1^{-1} equal to the mean distance between its occurrences, $\sum p_i n_i$. The second part is true since almost all blocks of large length N have approximated the proper frequency of each B_i .

Now we return to the construction of a subset of growth $2^{(H - \delta)N}$ all of equal probability. Let us choose integers

a_1 as close as possible to

$$p_1 N$$

and construct sequences with a_1 of the block B_1 . The number of blocks is then

$$\sum a_1 N = N \sum p_1 = N$$

and the number of sequences:

$$N \sum p_1 \log p_1$$

The growth is then in terms of symbols

$$\frac{\sum p_1 \log p_1}{\sum p_1 l_1} = P \sum p_1 \log p_1 = H$$

This proves the following:

Theorem: Given $\delta > 0$ there exists a set of M blocks of length N (when N is sufficiently large) such that

$$M > 2^{(S - \delta)N}$$

and each block has the same probability, and starts and ends in the same state, which can be chosen arbitrarily.

In case the system is not ergodic but made up of a finite number of ergodic systems:

$$r = \sum c_1 r_1$$

each r_1 will have a rate H_1 which we may assume arranged in a now increasing sequence

$$H_1 \geq H_2 \geq H_3 \geq \dots \geq H_n$$

The function $\varphi(\epsilon)$ then becomes a decreasing step function in the manner indicated by the following:

Theorem: In the case considered

$$\varphi(\epsilon) = H_1 \text{ in the interval } \frac{1}{2^{K-1}} < \epsilon < \frac{1}{2^{K'}}$$

For if ϵ is in the range indicated we must take a set of positive probabilities from at least one of r_1, \dots, r_1 . This gives a growth of type

$$2^{(H_k - \delta)N}$$

at least, and can be limited to this by choosing all sequences in r_k, r_{k+1}, \dots, r_n .

The quantity

$$H = 2H_1$$

will be called the mean statistical rate for the system.

C. E. SHANNON

April 26, 1948

[41]

Samples of Statistical English

C E Shannon

A number of samples of statistical English including probability structure out to four words are given below. These were constructed by starting off with three words from a book. These three words are shown to someone who fits them in a reasonable English sentence and writes down the word following the three. The first word is then covered up and the process repeated with a different person, etc. If the imagined sentence ends after the added word, the person writing the word adds a period. For samples bearing a title the participants were told that this was the subject dealt with. These samples may be compared with those in "A Mathematical Theory of Communication" where less statistical structure is included.

The samples given here were obtained, for the most part, with the aid of J. R. Pierce, B. McMillan, C. C. Cutler and W. E. Mathews. A few of the samples were obtained from other sources (contemporary literature, etc.) and are included for comparison. The reader may try his skill at guessing which are statistically constructed. The true sources are given at the end.

1. This was the first. The second time it happened without his approval. Nevertheless it cannot be done. It could hardly have been the only living veteran of the foreign power had stated that never more could happen. Consequently people seldom try it.
2. John now disported a fine new hat. I paid plenty for the food. When cooked asparagus has a delicious flavor suggesting apples. If anyone wants my wife or any other physicist would not believe my own eyes. I would believe my own word.
3. That was a relief whenever you be let your mind go free who knows if that pork chop I took with my cup of tea after was quite good with the heat I couldn't smell anything off it I'm sure that queer looking man in the
4. In a few days was the minimum amount of money remaining to the end. However everyone knows the meaning implied. It was true when Cutler says that we should proceed carefully. When you love yourself too much. The woman who accosted
5. Fourscore and twenty years passed before we could meet them that isn't already done should have been a good son is going fast according to the teacher of his ability. His intelligence sufficed for the time. This cannot change much.

6. Even the killing was atrociously perpetrated by the cruelest treatment that a small boy jumped over the hedge and buried her. A grave fault of many approaches to the furthestmost reaches of the state. Politics and business are becoming lost to the.
7. It is an Italian ox mouth dish. The only thing in the room is worms. I am the director of the seminar. In an evolving hemisphere. C'est Monsieur Jardin. I am a patient. Oh my dear Plapsen, you are my dearest Klapsen.
8. He took it with many other matters are more apparent if they think so. Is there a reason for supposing that most people don't. Nevertheless sex is absolutely necessary as though the electron diffraction camera plate up on the top surface of
9. Fifteen years before the mast, he ever had eaten. Try it and see. I believe that whatever arises a fund has been accumulated sufficiently in the near future holds many surprises. No man can judge his actions by his wife Susie.
10. I forget whether he went on and on. Finally he stipulated that this must stop immediately after this. The last time I saw him when she lived. It happened one frosty look of trees waving gracefully against the wall. You never can
11. When I bought my wife a long time ago. I knew that it wasn't faster when he didn't eat or drink a toast to John Doe, otherwise known as McMillan's theorem. Whatever the nature of Christ's teachings. Go far into
12.

McMillan's Theorem

McMillan's theorem states that whenever electrons diffuse in vacua. Conversely impurities of a cathode. No substitution of variables in the equation relating these quantities. Functions relating hypergeometric series with confluent terms converging to limits uniformly expanding rationally to represent any function.
13.

House Cleaning

First empty the furniture of the master bedroom and bath. Toilets are to be washed after polishing doorknobs the rest of the room. Washing windows semi-annually is to be taken by small aids such as husbands are prone to omit soap powder.

14. Epiminondas

Epiminondas was one who was powerful especially on land and sea. He was the leader of great fleet maneuvers and open sea battles against Pelopidas but had been struck on the head during the second Punic war because of the wreck of an armored frigate.
15. Salaries

Money isn't everything. However, we need considerably more incentive to produce efficiently. On the other hand too little and too late to suggest a raise without a reason for remuneration obviously less than they need although they really are extremely meager.
16. Murder Story

When I killed her I stabbed Claude between his powerful jaws clamped cruelly together. Screaming loudly despite fatal consequences in the struggle for life began ebbing as he coughed hallowly spitting blood from his ears. Burial seemed unnecessary since further division was necessary.

The sources are: 3, from "Ulysses" by James Joyce, page 748; 7 and 14 are the conversation and writings of two schizophrenic patients (quoted from Bleuler, "A Textbook of Psychiatry"). All others constructed by statistical means.

C. E. SHANNON

~~June 11, 1948~~

U42

[45]

DIGEST SERIES NO. 14
LCM 57/1 11-13 October 1948

SYMPOSIUM ON COMMUNICATION RESEARCH

The Department of Defense
RESEARCH AND DEVELOPMENT BOARD
Washington 25, D. C.

Prepared by
THE PANEL OF COMMUNICATIONS OF
THE COMMITTEE ON ELECTRONICS

Approved: K. C. Black
Chairman

5. SIGNIFICANCE AND APPLICATION

C. E. Shannon
Bell Telephone Laboratories
Murray Hill, N. J.

1. Introduction.

A general communication system is shown in Figure 3. An information source produces a message. This is encoded in a transmitter to produce a signal suitable for transmission over the channel. During transmission the signal may be perturbed by noise. The perturbed signal is decoded or demodulated at the receiver to recover, as well as possible, the original message.

The situation is roughly analogous to a transportation system for transporting physical goods from one point to another. We can imagine, for example, a lumber mill producing lumber at an average rate of R cubic feet per second and a conveyor system capable of transporting C cubic feet per second. If R is greater than C the full output of the mill cannot possibly be carried on the conveyor. On the other hand, if R is less than or equal to C it may or may not be possible, depending on whether the lumber can be efficiently packed in the available space of the conveyor. However, if we allow ourselves to saw the lumber up into suitable sizes and shapes we can always approach 100 per cent efficiency in packing. In this case we must, of course, supply a carpenter shop at the other end of the conveyor to reassemble the lumber in its original form before passing it on to the consumer.

If the analogy is sound we might hope to define two parameters R and C associated with an information source and a channel, respectively. R should measure, in some sense, how much information is produced per second by the source, and C the capacity of the channel when used in the most efficient manner for transmitting information. We would expect then that if $R > C$ the full output of the source cannot be transmitted satisfactorily. If $R \leq C$ it should be possible to transmit the output of the source by proper encoding and decoding at transmitter and receiver. It turns out that it is possible to define quantities R and C which measure these information rates and capacities and satisfy the desired relationships. We will attempt to show how this can be done without, however, giving mathematical proofs of the results.¹

2. The Information Source.

The first problem is that of clarifying the nature of "information" and finding a measure of the rate of production for an information source.

Information involves basically the concept of "choice." An information source chooses one particular message from a set of possible messages. If there were only

¹For mathematical details, see Shannon, C.E., "A Mathematical Theory of Communication," Bell System Technical Journal, July and October, 1948. See also Shannon, C.E., "Communication in the Presence of Noise," Proceedings of the I.R.E. (Forthcoming).

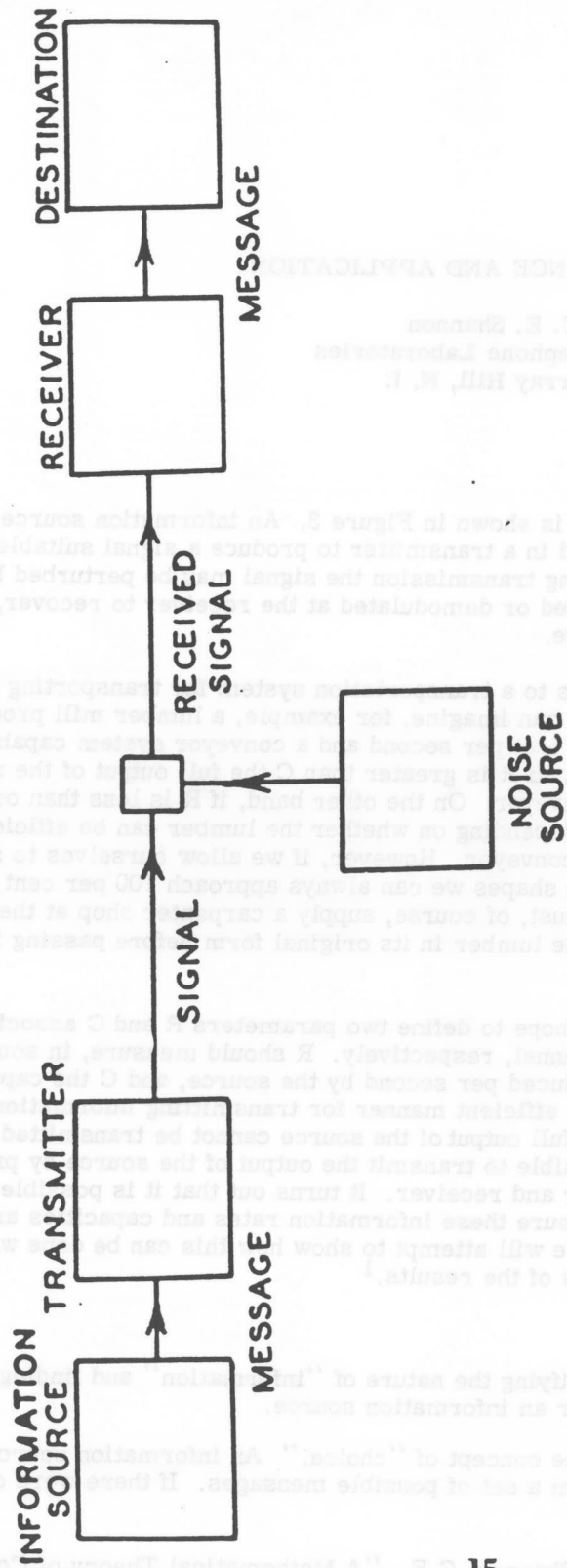


Figure 3

one possible message there would be no communication problem. The amount of information produced by a source must evidently be related to the range of choice available.

The simplest possible choice is a choice from two equally likely possibilities, say 0 or 1. We shall call the corresponding unit of information a binary digit or "bit." A relay or flip-flop circuit has two possible states and is capable of storing one bit of information.

A device which chooses at random from 0 or 1 making one choice each second is considered to be producing information at rate R of one bit per second. Such a source produces a "message" which is a random sequence of 0's and 1's.

A choice from say 32 equally likely possibilities can be considered as a series of five choices, each from two equally likely possibilities, and, therefore, should correspond to five bits. More generally, a choice from n equally likely possibilities represent $\log_2 n$ bits.

Suppose now that the various possible choices have different probabilities of occurrence, say p_1, p_2, \dots, p_n . How much information is produced when a choice is made under these circumstances? One feels intuitively that less "choice" is involved in a device which chooses between 0 and 1 with probabilities .01 and .99 than in one which chooses with equal probabilities. In the former case the result is almost sure to be 1.

The following example shows that by proper encoding an average compression can be obtained by using the probabilities p_1, p_2, \dots, p_n . Suppose there are four possible choices A, B, C, D with probabilities $p_A = 1/2, p_B = 1/4, p_C = 1/8, p_D = 1/8$. If we use a simple direct code into binary digits:

A = 00 B = 01 C = 10 D = 11,

we use two binary digits per letter. On the other hand, using the following code where more probable letters are given short codes and less probable letters longer codes, we obtain an average saving

A = 0 B = 10 C = 110 D = 111.

This is a reversible code; the original text can be recovered from the encoded sequences as is readily verified. With this code we need, on the average, only

$$(1/2 \times 1 + 1/4 \times 2 + 1/8 \times 3 + 1/8 \times 3) = 1 \frac{3}{4}$$

binary digits per letter. We may say then that a choice with probabilities $1/2, 1/4, 1/8, 1/8$ corresponds to $1 \frac{3}{4}$ bits of information. If an information source were producing a sequence of the letters A, B, C, D with these probabilities we could encode it into a sequence of binary digits in which $1 \frac{3}{4}$ binary digits are used on the average for each letter of message.

A general analysis of the situation shows that if the letters are chosen with probabilities p_1, p_2, \dots, p_n then it is possible to encode into binary digits using

$$H = - \sum p_i \log_2 p_i$$

binary digits per letter of message on the average, and there is no method of reversible encoding using less. This H then is the equivalent number of bits per letter, and, if the source produces n letters per second, $R = nH$ is the rate of production in bits per second.

In the case of English text the statistical structure is more involved. There are the various letter probabilities p_i , but, also, there are statistical influences between nearby letters. For example, the letter T is more often followed by H than by any other letter, Q is almost invariably followed by U, etc. In such cases there is a more general formula for calculating the equivalent number of bits per letter of message. Let $p(i, j, \dots, s)$ be the probability in the language of the sequence of letters i, j, \dots, s . Then we define G_n by

$$G_n = -\frac{1}{n} \sum p(i, j, \dots, s) \log_2 p(i, j, \dots, s)$$

where the sum is over all sequences of letters which are just n letters long. The sequences $G_1, G_2, \dots, G_n, \dots$ represents a series of approximations to the desired H which takes into account more and more of the statistical structure as we proceed along the sequence. The information per letter of message can be defined by the limiting value of the G 's.

$$H = \lim_{n \rightarrow \infty} G_n$$

It can be shown that H has the desired properties; namely, we can encode the messages from the source into binary digits using H binary digits per letter on the average, and no method of encoding uses less.

For the English language H has been estimated at roughly 2 bits per letter, taking account only of the statistical structure out to about 6 or 8 letters.

If the messages produced by the information source are continuous functions of time, as in speech or television transmission, the situation is much more involved and we will not discuss it in detail. It is still possible to assign a rate of production of information in bits per second to such a source, but the rate now depends on other considerations. With continuous functions as messages, exact reproduction is not generally required, and the rate R depends on the amount and nature of the discrepancy which can be tolerated between the original and recovered messages. The tolerable discrepancy in turn is determined by the final destination of the messages. With speech, for example, the tolerable errors depend on the structure of the human ear and brain.

Although the mathematical problems involved in defining the rate for a continuous source have been completely solved, it is in practical cases very difficult to estimate R . The following calculation may be of some interest, however. Suppose we are interested only in transmitting English speech (no music or other sounds), and the quality requirements on reproduction are only that it be intelligible as to meaning. Personal accents, inflections, etc., can be lost in the process of transmission. In such a case we could, at least in principle, transmit by the following scheme. A device is constructed at the transmitter which prints the English text corresponding to the spoken words. These can be translated into binary digits in the ratio of about two binary digits per letter, or $2 \times 4.5 = 9$ per word. Taking 100 words per minute as a reasonable talking speed we obtain 900 bits per minute or 15 bits per second as an estimate of the rate for English speech when intelligibility is the only fidelity requirement.

3. The Capacity of a Channel.

We now consider the problem of defining the capacity C of a channel for transmitting information. Since we have measured the rate of production for an information source in

bits per second, we would naturally like to measure C in the same units. The question then becomes, what is the maximum number of binary digits per second that can be transmitted over a given channel?

In some cases the answer is simple. With a teletype channel there are 32 possible symbols. Each symbol therefore represents 5 bits. If we can send n symbols per second, and the noise level is not high enough to introduce any errors during transmission, we can send $5n$ bits per second.

Suppose now that the channel is defined as follows. We can use for signals any functions of time $f(t)$ which lie within a certain band of frequencies, W cycles per second wide. It is known that a function of this type can be specified by giving its value at a series of equally spaced sampling points $\frac{1}{2W}$ seconds apart. Thus we may say that such a function has $2W$ degrees of freedom, or dimensions, per second.

If there is no noise whatever on such a channel we can distinguish an infinite number of different amplitude levels for each sample. Consequently we could, in principle, transmit an infinite number of binary digits per second, and the capacity C would be infinite.

Even when there is noise, if we place no limitations on the transmitter power the capacity will be infinite, for we may still distinguish at each sample point an unlimited number of different amplitude levels. Only when noise is present and the transmitter power is limited in some way do we obtain a finite capacity C . The capacity depends, of course, on the statistical structure of the noise as well as the nature of the power limitation.

The simplest type of noise is white thermal noise or resistance noise. The probability distribution of amplitudes is Gaussian, and the spectrum is flat with frequency over the band in question and may be assumed to be zero outside the band. This type of noise is completely specified by giving its mean square amplitude N the power it would deliver into a unit resistance.

The simplest limitation on transmitter power is to assume that the average power delivered by the transmitter (or more precisely the mean square amplitude of the signal) be not more than P . If we define our channel by these three parameters W , P , and N , the capacity C can be calculated. It turns out to be

$$C = W \log_2 \frac{P + N}{N} \text{ (bits per second).}$$

It is easy to see that this formula is approximately right when P/N is large. The received signal will have a power $P + N$, and we can distinguish something of the order of

$$\sqrt{\frac{P + N}{N}}$$

different amplitudes at each sample point. In a time T there will be $2TW$ independent samples. Thus, there are approximately

$$M = \left(\sqrt{\frac{P + N}{N}} \right)^{2TW} = \left(\frac{P + N}{N} \right)^{TW}$$

different signal functions of duration T that can be distinguished from one another in spite of the noise. This corresponds to

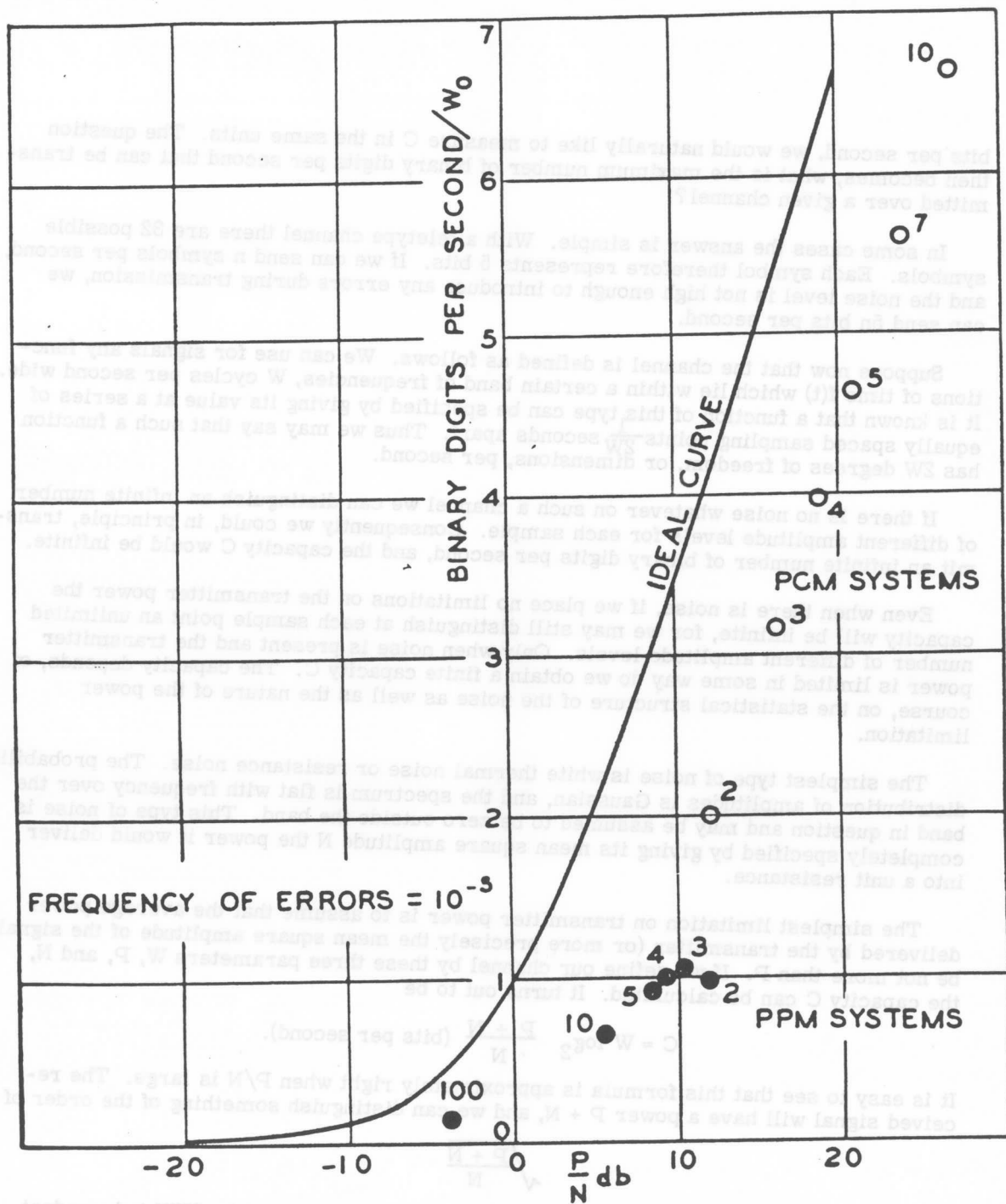


Figure 4

$$\log_2 M = TW \log_2 \frac{P+N}{N}$$

binary digits in the time T or

$$C = W \log_2 \frac{P+N}{N}$$

binary digits per second. This formula has a much deeper and more precise significance than the above argument would indicate. In fact it can be shown that it is possible, by properly choosing our signal functions, to transmit $W \log_2 \frac{P+N}{N}$ binary digits per second with as small a frequency of errors as desired. It is not possible to transmit binary digits at any higher rate with an arbitrarily small frequency of errors. This means that the capacity is a sharply defined quantity in spite of the noise. These statements are proved by two different methods.²

The formula for C applies for all values of P/N. Even when P/N is very small, the average noise power being much greater than the average transmitter power, it is possible to transmit binary digits at the rate $W \log_2 \frac{P+N}{N}$ with as small a frequency of errors as desired. In this case $\log_2 (1 + \frac{P}{N})$ is approximated by $\frac{P}{N} \log_2 e = 1.443 \frac{P}{N}$ and we have approximately

$$C = 1.443 \frac{PW}{N}$$

It should be emphasized that it is only possible to transmit at a rate C over a channel by properly encoding the information. In general, the rate C is only approached as a limit by using more and more complex encoding and longer and longer delays at both transmitter and receiver. In the white noise case the best encoding is such that the transmitted signals themselves have the structure of a white noise with power P. The difficulty with the approximate argument given for that case, and the reason it does not give a sharply defined capacity, is that the selection of signals is not optional. The distribution of amplitudes is not Gaussian as it should be.

4. Comparison of Ideal and Practical Systems.

In Figure 4 the curve is the function

$$\frac{C}{W} = \log (1 + \frac{P}{N})$$

plotted against P/N measured in db. It represents, therefore, the channel capacity per unit of band with white noise. The circle and points correspond to PCM and PPM systems used to send a sequence of binary digits and adjusted to give about one error in 10^4 binary digits. In the PCM case the number adjacent to a point represents the number of amplitude levels - 3 for example is a ternary PCM system. In all cases positive and negative amplitudes are used. The PPM systems are quantized with a discrete set of possible positions for the pulse, the spacing is $\frac{1}{2W}$, and the number adjacent to a point is the number of possible positions for a pulse.

The series of points follows a curve of the same shape as the ideal but displaced horizontally about 8 db. This means that with more involved encoding or modulation systems a gain of 8 db. in power could be achieved over the system indicated.

²See Shannon, C. E., "Mathematical Theory of Communication" and "Communication in the Presence of Noise."

Of course, as one attempts to approach the ideal, the transmitter and receiver required become more complicated and the delays increase. For these reasons there will be some point where an economic balance is established between the various factors. It may well be, however, that even at the present time more complex systems would be justified.

A curious fact illustrating the general misanthropic behaviour of Nature is that at both extremes of P/N (when we are well outside the practical range) the series of points in Figure 4 approach more closely the ideal curve. At very large P/N the PCM points approach to within $10 \log_{10} \frac{11e}{3} = 4.5$ db. of the ideal while with very small P/N the PPM points approach to within 3 db. The relation

$$C = W \log \left(1 + \frac{P}{N} \right)$$

can be regarded as an exchange relation between the parameters W and P/N. Keeping the channel capacity fixed we can decrease the bandwidth W provided we increase P/N sufficiently. Conversely, an increase in band allows a lower signal-to-noise ratio in the channel. The required P/N in db. is shown in Figure 5 as a function of the band W. It is assumed here that as we increase W, N increases proportionally:

$$N = W N_0$$

where N_0 is the noise power per cycle of band. It will be noticed that if P/N is large a reduction of band is very expensive in power. Halving the band roughly doubles the signal-to-noise ratio in db. that is required.

The channel capacity C can be calculated in many other cases. A general result that applies in any situation where the average transmitter power is limited to P is that the channel capacity is bounded by:

$$W \log \frac{P + N_1}{N_1} \leq C \leq W \log \frac{P + N}{N_1}$$

where N_1 is a parameter called the "entropy power" of the noise. It is defined as the power in a white noise having the same entropy as the actual noise. N is, as before, the average noise power.

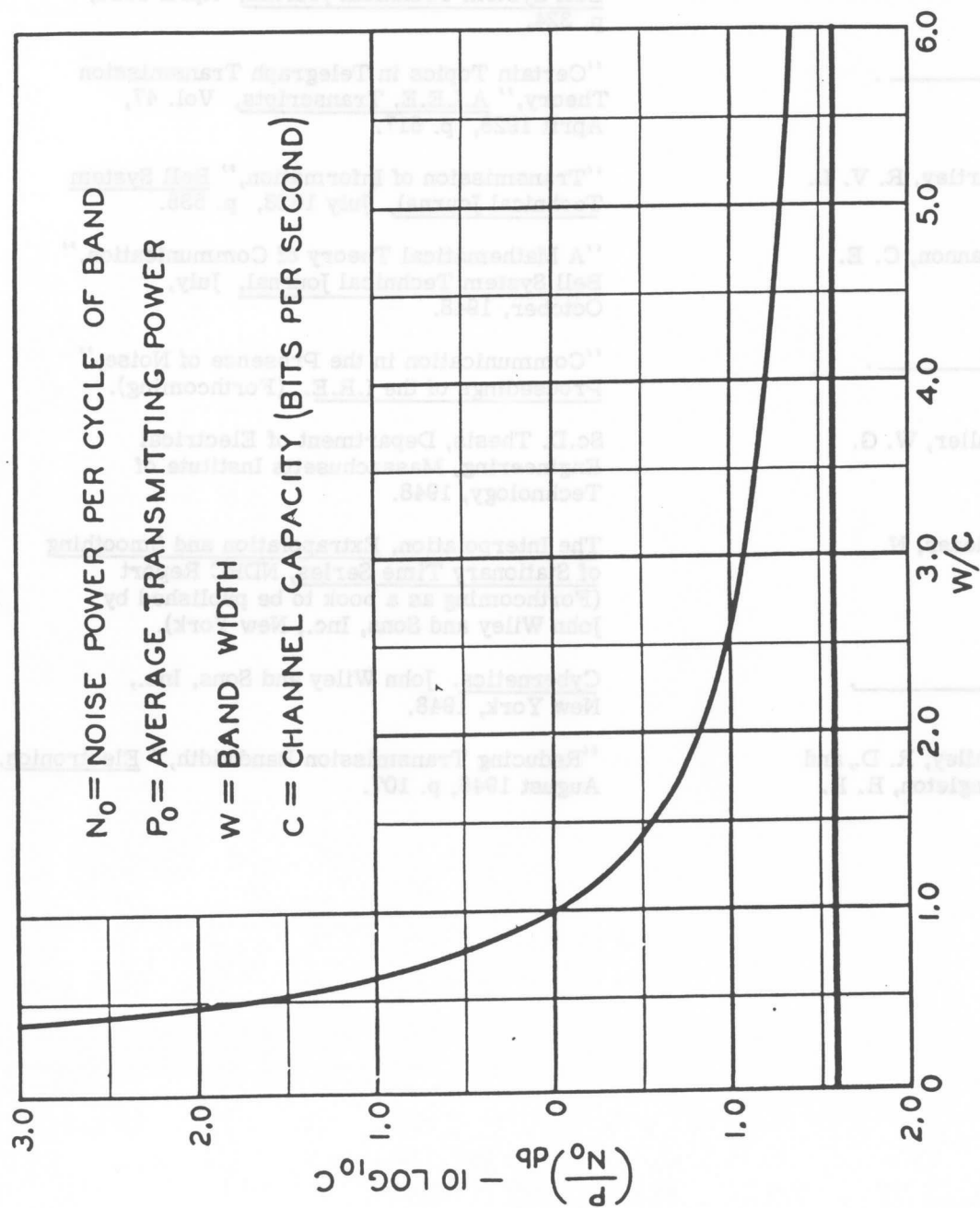


Figure 5

REFERENCES

- Nyquist, H. "Certain Factors Affecting Telegraph Speed," Bell System Technical Journal, April 1924, p. 324.
- _____. "Certain Topics in Telegraph Transmission Theory," A.I.E.E. Transcripts, Vol. 47, April 1928, p. 617.
- Hartley, R. V. L. "Transmission of Information," Bell System Technical Journal, July 1928, p. 535.
- Shannon, C. E. "A Mathematical Theory of Communication," Bell System Technical Journal, July, October, 1948.
- _____. "Communication in the Presence of Noise," Proceedings of the I.R.E. (Forthcoming).
- Tuller, W. G. Sc.D. Thesis, Department of Electrical Engineering, Massachusetts Institute of Technology, 1948.
- Wiener, N. The Interpolation, Extrapolation and Smoothing of Stationary Time Series, NDRC Report (Forthcoming as a book to be published by John Wiley and Sons, Inc., New York).
- _____. Cybernetics. John Wiley and Sons, Inc., New York, 1948.
- Balley, R. D., and Singleton, H. E. "Reducing Transmission Bandwidth," Electronics, August 1948, p. 107.

[46]

Note on Certain Transcendental Numbers

Claude E. Shannon

This note calls attention to a certain class of numbers that are easily shown to be transcendental but seem to have escaped previous notice. A typical example is the number

$$\lambda = 2^{-2^{-2^{\cdot^{\cdot^{\cdot}}}}},$$

or more precisely $\lambda = \lim_{n \rightarrow \infty} \lambda_n$, $\lambda_{n+1} = 2^{-\lambda_n}$, $\lambda_0 = 2$. It is easily seen that λ exists and satisfies the equation $\lambda = 2^{-\lambda}$. It is known from a conjecture of Hilbert, proved by Gelfond and by Schneider, that a^x is transcendental if $a \neq 0, 1$ is algebraic and x is an algebraic irrational. Now λ is clearly not rational, and if we suppose it an algebraic irrational, it must then be transcendental, a contradiction. Hence it is transcendental.

More generally let f be a function such that if x is algebraic and does not belong to a set S , then $f(x)$ is transcendental. Let g_1 and g_2 be algebraic functions and such that $x \notin g_1 f g_2 x$, $x \in S$. Then the solutions of

$$x = g_1 f g_2 x$$

are transcendental by a similar argument, using the fact that g_1^{-1} is algebraic. If the sequence $\lambda_n = (g_1 f g_2)^n \lambda_0$ approaches a limit λ it must be transcendental. Some functions known to have the property required for f are $\sin x$, e^x and $J_0(x)$, the exceptional set S consisting of the number 0.

C. E. SHANNON

October 27, 1948

446

[47]

A CASE OF EFFICIENT CODING FOR A VERY NOISY CHANNEL

Consider a discrete channel with two possible symbols 0 and 1. Noise is assumed to affect successive symbols independently and in such a way that the probability of a symbol being interpreted correctly at the receiver is $p = \frac{1+\epsilon}{2}$ which is

the probability of incorrect interpretation is $q = \frac{1-\epsilon}{2}$.

The capacity of such a channel is

$$\begin{aligned} C &= \text{Max } H(y) = H_x(y) \\ &= -\log 2 + \frac{1+\epsilon}{2} \log \frac{1+\epsilon}{2} + \frac{1-\epsilon}{2} \log \frac{1-\epsilon}{2} \\ &= -\log 2 + \frac{1+\epsilon}{2} \log (1+\epsilon) + \frac{1-\epsilon}{2} \log (1-\epsilon) - \log 2 \\ &= \frac{1+\epsilon}{2} \log (1+\epsilon) + \frac{1-\epsilon}{2} \log (1-\epsilon) \end{aligned}$$

We assume ϵ very small and approximate $\log (1 \pm \epsilon)$ by $\pm \epsilon \mp \frac{\epsilon^2}{2}$.

$$\begin{aligned} C &\approx \frac{1+\epsilon}{2} \left(\epsilon - \frac{\epsilon^2}{2} \right) + \frac{1-\epsilon}{2} \left(-\epsilon - \frac{\epsilon^2}{2} \right) \\ &= \epsilon^2 - \frac{\epsilon^3}{2} \\ &\approx \epsilon^2 \text{ (natural units)} \end{aligned}$$

In bits per symbol, the capacity is:

$$C = \epsilon^2 \log_2 e$$

A very simple code can be constructed for this system to send a sequence of random binary digits at nearly the rate C with a quite small frequency of errors; in other words a code which is not far from the ideal. The code is merely to repeat each binary digit in the message a large number n of times. At the receiver, a group of n is received, and the majority report is taken as the original message symbol.

If the message symbol is 0 then n 0's are transmitted. At the receiver the n received symbols will be a mixture of 0's and 1's the number of 0's present will be distributed according to a binomial distribution with $p = \frac{1+\epsilon}{2}$ and $q = \frac{1-\epsilon}{2}$.

For large n the binomial distribution is approximately normal (and this approximation is especially good when p is close to $\frac{1}{2}$). The expected number of 0's is $p n$, and the standard deviation is:

$$\sigma = \sqrt{n p q} = \sqrt{\frac{n(1-\epsilon^2)}{4}}$$

An error occurs when the number of received 0's is less than $\frac{n}{2}$, i.e. when the actual number of zeros is $p n - \frac{n}{2}$ away from the expected number. In terms of σ this is:

$$a = \frac{p n - \frac{n}{2}}{\sqrt{\frac{n(1-\epsilon^2)}{4}}} = \frac{\epsilon \sqrt{n}}{\sqrt{1-\epsilon^2}} \text{ standard deviations.}$$

Hence the frequency of errors is given by the area of a normal curve with standard deviation equal to unity from a out to ∞ .

To obtain a frequency of errors 10^{-3} , say, we must have $a = 1.5$

$$n = \frac{2.3}{\epsilon^2}$$

and the rate is $\frac{\epsilon^2}{2.3}$ as compared with the rate $1.45 \epsilon^2$ for the ideal (with essentially zero frequency of errors).

C. E. SHANNON

November 18, 1948

[48]

December 6, 1948

Note on Reversing A Discrete Markhoff Process

In "A Mathematical Theory of Communication" a language was represented by a discrete Markhoff process with a finite number of possible states. Such a stochastic process can be represented schematically by means of an oriented linear graph as in Fig. 1

Consider the question of generating the same language in reverse; for example, English but read backwards. Can we always invert a finite state Markhoff process and obtain a finite state Markhoff process? The answer is "yes" and furthermore the corresponding linear graph has the same topology, but with reversed ~~inversed~~ orientation on all branches. If the original process has ^{state} ~~probabilities~~ P_i and transition probabilities P_{ij} (probability when in state i of going to state j), then the reverse process has the same state probabilities and the transition probabilities given by:

$$q_j(i) = \frac{P_i}{P_j} P_{ij}(j)$$

This is true since this $q_j(i)$ is merely the a posteriori probability for the original process that when in state j the preceding state was state i . The inverse of Fig. 1 is shown in Fig. 2.

It is interesting to show directly that the entropy H_R of the reverse process is equal to the entropy H_F of the forward process. Of course, this must be true a posteriori from the general properties of entropy. We have

$$P_j q_j(i) = P_i P_{ij}(j)$$

Hence:

$$\sum P_j q_j(i) \log P_j q_j(i) = \sum P_i \cancel{P_i(j)} \log P_i \cancel{P_i(j)}$$

or

$$\begin{aligned} \sum P_j q_j(i) \log q_j(r) + \sum P_j q_j(i) \log P_i \\ = \sum P_i \cancel{P_i(j)} \log \cancel{P_i(j)} + \sum P_i \cancel{P_i(j)} \log P_i \end{aligned}$$

Hence:

$$-H_R + \sum P_j \log P_j = -H_F + \sum P_i \log P_i$$

$$H_R = H_F$$

C. E. SHANNON

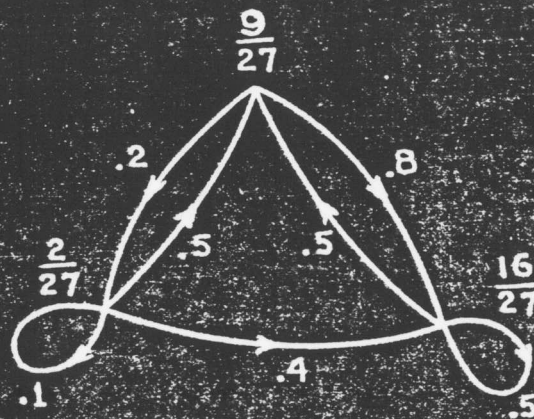


Fig. 1

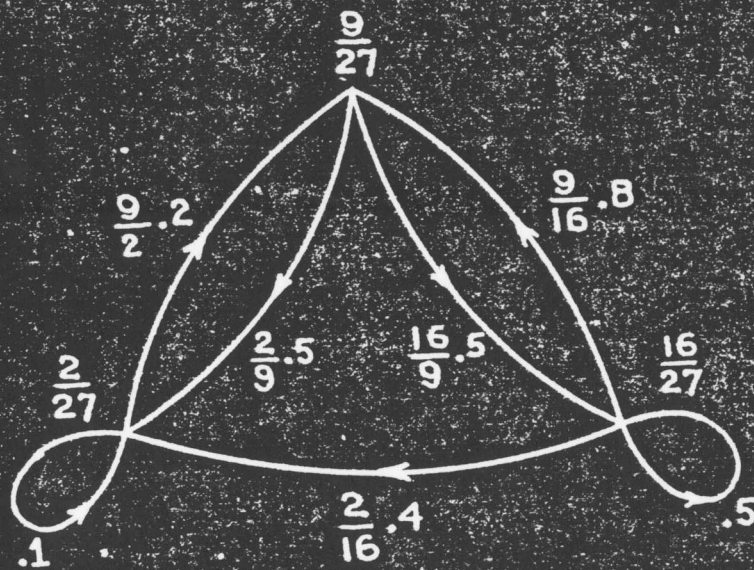


Fig. 2

Outline of Talk

American Statistical Society, December 28, 1949

INFORMATION THEORY

by

C. E. Shannon

Bell Telephone Laboratories, Inc., Murray Hill, N. J.

1. Information Produced by a Stochastic Process

In communication engineering, we are interested in transmitting messages from one point to another. The messages generally consist of a sequence of individual symbols, such as the letters of printed English, which are governed by probabilities. Thus, in English, there are the various letter frequencies, digram frequencies, etc. The "meaning" of the message (if any) is irrelevant to the engineering problem. Abstractly, then, we may consider a message to be a sequence of meaningless symbols produced by a suitable Stochastic process. Communication systems must be designed to handle the ensemble of possible messages; the particular one which will actually occur is not known when the system is constructed. The source producing messages is assumed to have only a finite number of possible internal states.

2. Entropy as a Measure of Information

A suitable measure of the amount of information produced by a discrete Stochastic process is given by the entropy H , where

$$H = - \lim_{N \rightarrow \infty} \frac{1}{N} \sum p(x_1, \dots, x_N) \log_2 p(x_1, \dots, x_N)$$

in which x_1, \dots, x_N is a sequence of N symbols produced by the process, $p(x_1, \dots, x_N)$ is the probability of this sequence, and the sum is over all sequences of this length.

The significance of the quantity H is that it is possible to translate messages from a source with entropy H into a sequence of binary digits (0 or 1) using, on the average, $H + \epsilon$ binary digits per letter of the original message with any positive ϵ . It is not possible to translate so that fewer are used. Thus, H measures, in a sense, the equivalent number of binary digits per letter of message. It can be shown that H also determines the amount of channel capacity required for transmission of the original messages.

Another important concept is that of the conditional entropy, $H_x(y)$, of one source relative to another. This measures in a sense the uncertainty per letter of the y sequence when the x sequence is known, or the amount of additional information in the y sequence over that available in the x sequence. $H_x(y)$ can be defined as follows:

$$H_x(y) = H(x, y) - H(x)$$

where $H(x, y)$ is the entropy of the sequence whose elements are the ordered pairs (x, y) .

3. The Nature of Information

While the entropy H measures the amount of information produced by a Stochastic process, it does not define the information itself. Thus two entirely difference sources might

produce information at the same rate (same H) but certainly they are not producing the same information. If we translate the output of a particular source into a different "language" by a reversible operation, the translation may be said to have the same information as the original. Thus we are led to consider the information of a Stochastic process as that which is common to all translations obtained from the given process by members of the group G of reversible translations, or, alternatively, as the equivalence class of all processes obtained from the given one by such translations. To avoid certain paradoxical situations, involving infinite internal storage in the transducer doing the translating, it is desirable to first limit the group G to translations possible in transducers having a finite number of possible internal states. The information associated with a process may be denoted by a single letter, say X . Thus $X = Y$ means that Y can be obtained by a translation of X , and conversely. It is possible to set up a metric satisfying the usual postulates as follows:

$$\begin{aligned} \rho(x, y) &= H_x(y) + H_y(x) \\ &= 2H(x, y) - H(x) - H(y) . \end{aligned}$$

With this metric it is possible to define limiting sequences of elements, each of which is an information. Thus a Cauchy sequence, X_1, X_2, \dots , is defined by requiring that

$$\lim_{\substack{m \rightarrow \infty \\ n \rightarrow \infty}} \rho(X_m, X_n) = 0 .$$

The introduction of these sequences as new elements (analogous to irrational numbers) completes the space in a satisfactory way and enables one to simplify the statement of various results.

4. The Information Lattice

A relation of inclusion, $x \geq y$, between two information elements x and y can be defined by

$$x \geq y \Leftrightarrow H_x(y) = 0.$$

This essentially requires that y can be obtained by a suitable finite state operation (or limit of such operations) on x . If $x \geq y$ we call y an abstraction of x . If $x \geq y$, $y \geq z$, then $x \geq z$. If $x \geq y$, then $H(x) \geq H(y)$. Also $x > y$ means $x \geq y$, $x \neq y$. The information element, one of whose translations is the process which always produces the same symbol, is the 0 element, and $x \geq 0$ for any x .

The sum of two information elements, $z = x + y$, is the process which produces the ordered pairs (x_n, y_n) . We have

$$z \geq x, \quad z \geq y$$

and there is no $u < z$ with the properties; z is the least upper bound of x and y .

The product $z = xy$ is defined as the largest z such that $z \geq x$, $z \geq y$; that is, there is no $u > z$ having both x and y as abstractions. The product is unique.

With these definitions information elements form a metric lattice. The lattice is not distributive, nor even modular. A non-distributive example is x, y independent sequences of binary digits, with z the sequence obtained by mod 2 addition of corresponding symbols in x and y . Then

$$zy + zx = 0 + 0 = 0$$

$$z(x + y) = z \neq 0.$$

The lattices are relatively complimented. There exists for $x \leq y$ a z with

$$z + x = y$$

$$zx = 0.$$

The element z is not, in general, unique.

5. The Delay Free Group G_1

The definition of equality for information based on the group G allows $x = y$ when y is, for example, a delayed version of x ; $y_n = x_{n+a}$. In some situations, when one must act on information at a certain time, a delay is not permissible. In such a case we may consider the more restricted group G_1 of instantaneously reversible translations. One may define inclusion, sum, product, etc., in an analogous way, and this also leads to a lattice but of much greater complexity and with many different invariants.

Proof of an Integration Formula

C. E. Shannon

The integral

$$f_N(\alpha) = \int_0^\alpha \frac{\sin^2 N x}{\sin^2 x} dx = -\frac{1}{2} \int_0^\alpha \frac{\cos 2 N x - 1}{\sin^2 x} dx \quad (1)$$

has arisen in an acoustical problem. It has been evaluated for $N = 1, 2, 3, 4$ as equal to

$$g_N(\alpha) = \alpha N + \sum_{i=1}^{N-1} \frac{N-i}{i} \sin 2 i \alpha \quad (2)$$

by R. C. Jones, and he has conjectured that $f_N = g_N$ for all α, N . A general proof follows.

From (1) we have

$$\Delta_{N,N}^2 f_N(\alpha) = f_N - 2f_{N-1} + f_{N-2} = -\frac{1}{2} \int_0^\alpha \frac{\cos 2Nx - 2\cos 2(N-1)x + \cos 2(N-2)x}{\sin^2 x} dx$$

and

$$\frac{d}{d\alpha} \Delta_{N,N}^2 f_N(\alpha) = -\frac{\cos 2Na - 2\cos 2(N-1)\alpha + \cos 2(N-2)\alpha}{2\sin^2 \alpha} \quad (3)$$

Also from (2)

$$\Delta_N g_N(\alpha) = \alpha + \sum_{i=1}^{N-1} \frac{\sin 2 i \alpha}{i}$$

$$\Delta_{N,N}^2 g_N(\alpha) = \frac{\sin 2(N-1)\alpha}{N-1}$$

$$\frac{d}{d\alpha} \Delta_{N,N}^2 g_N(\alpha) = 2\cos 2(N-1)\alpha \quad (4)$$

The equality of (3) and (4) can be established by noting that the numerator of (3),

$$\begin{aligned}
 & \cos 2N\alpha - 2 \cos 2(N-1)\alpha + \cos 2(N-2)\alpha \\
 &= \operatorname{Re} \left(e^{j2N\alpha} - 2 e^{j2(N-1)\alpha} + e^{j2(N-2)\alpha} \right) \\
 &= \operatorname{Re} \left(e^{j2(N-1)\alpha} \left[e^{j2\alpha} - 2 + e^{-j2\alpha} \right] \right) \\
 &= \operatorname{Re} \left(e^{j2(N-1)\alpha} (2j)^2 \left[\frac{e^{j\alpha} - e^{-j\alpha}}{2j} \right]^2 \right) \\
 &= - \operatorname{Re} \left\{ 4 \sin^2 \alpha e^{j2(N-1)\alpha} \right\} = -4 \sin^2 \alpha \cos 2(N-1)\alpha .
 \end{aligned}$$

Hence

$$\frac{d}{d\alpha} \Delta^2 g = \frac{d}{d\alpha} \Delta^2 f$$

but $\Delta^2 g_N(0) = \Delta^2 f_N(0) = 0$, so that

$$\Delta_{N,N}^2 g_N(\alpha) = \Delta_{N,N}^2 f_N(\alpha)$$

also it has been verified that

$$g_1(\alpha) = f_1(\alpha)$$

$$g_2(\alpha) = f_2(\alpha)$$

Hence it follows in general that

$$f_N(\alpha) = g_N(\alpha)$$

[59]

Fundamental Theory

It is possible by various methods of modulation to improve one aspect of a system for transmitting information at the expense of others. The various quantities which may be exchanged are:

1. quality of received signal, which can be roughly measured in some cases by the signal to noise ratio.
2. Transmitter power.
3. Time of transmission.
4. Band width of transmitted signal.
5. Noise in the system.

The general theory of how these variables are related and the amount of compression that is theoretically possible is quite involved and will be developed in a forthcoming memorandum.

However speaking roughly and under a number of auxiliary assumptions one can say that rates of exchange are controlled by the following equation:

$$D = f(2W \ln \frac{P}{N})$$

in which

D = a measure of distortion at the receiver

T = time of transmission

W = band width of transmitter

P = transmitter power

N = noise power density, that is the noise power per unit band width, the noise spectrum being assumed flat in the region under consideration

This means that keeping received quality invariant we may exchange T , B , P and N in various way so long as we keep the argument of the function,

$$T B \ln \frac{P}{N B}$$

the same. This may also be written

$$T B \ln \frac{P}{E_T B} = T B \ln \frac{E_T}{E_N}$$

where E_T and E_N are the total transmitter energy and noise energy during the transmission time. Thus for example by increasing band width we can decrease transmitter energy - this exchange is in one sense very favorable since it is a logarithmic one; adding units of band width divides the energy by a factor.

Frequency modulation and pulse position modulation are two methods of buying signal to noise ratio at the expense of band width. Neither of these however is by any means optimal in the exchange. The present memorandum describes a new method of modulation in which essentially the maximum saving of signal power is achieved for a given band width increase. This does not mean that the system of transmission is a theoretically ideal one for there are several other means of improving received quality keeping T , B , P and N fixed - what this system does is to yield a nearly ideal exchange rate between the variables.

The system is based on the idea of sending the values of the input modulating function (the speech function in telephone and radio) at a sequence of regularly spaced sampling

Thus $9 = 8 + 4 - 2 - 1$

and $-5 = -8 + 4 - 2 + 1$

A transmitter for this system could be built in the following way. A condenser is charged as usual to the sampled voltage. This voltage is read on a comparator biased up to half the maximum. If the comparator gives a positive indication an electronic switch is closed feeding a negative pulse of 2^n units of charge into the condenser; if not a positive pulse of 2^n units is fed in. The comparator is now switched to control a new pulse source which produces pulses of 2^{n-1} units and the process is repeated. Thus the circuit feeds in positive or negative pulses of decreasing magnitude "hunting" for a balance. At each stage a recorder remembers whether a positive or negative pulse was used. These positive and negative recordings actually are the binary representation of the original voltage, as one can see by reading the above table with 1' replaced by 0. Hence the receiver of Fig. 4 can be used without alteration in this system.

Creative Thinking

- Claude Shannon

Mar. 20, 1945

Up to 100% of the amount of ideas produced, useful good ideas produced by these signals, these are supposed to be arranged in order of increasing ability. At producing ideas, we find a curve something like this. Consider the number of curves produced here - going up to enormous height here.

A very small percentage of the population produces the greatest proportion of the important ideas. This is akin to an idea presented by an English mathematician, Turig, that the human brain is something like a piece of uranium. The human brain, if it is below the critical lap and you shoot one neutron into it, additional more would be produced by impact. It leads to an extremely explosive of the issue, increase the size of the uranium. Turig says this is something like ideas in the human brain. There are some people if you shoot one idea into the brain, you will get a half an idea out. There are other people who are beyond this point at which they produce two ideas for each idea sent in. Those are the people beyond the knee of the curve. I don't want to sound egotistical here, I don't think that I am beyond the knee of this curve and I don't know anyone who is. I do know some people that were. I think, for example, that anyone will agree that Isaac Newton would be well on the top of this curve. When you think that at the age of 25 he had produced enough science, physics and mathematics to make 10 or 20 men famous - he produced binomial theorem, differential and integral calculus, laws of gravitation, laws of motion, decomposition of white light, and so on. Now what is it that shoots one up to this

part of the curve? What are the basic requirements? I think we could set down three things that are fairly necessary for scientific research or for any sort of inventing or mathematics or physics or anything along that line. I don't think a person can get along without any one of these three.

The first one is obvious - training and experience. You don't expect a lawyer, however bright he may be, to give you a new theory of physics these days or mathematics or engineering.

The second thing is a certain amount of intelligence or talent. In other words, ^{you have} to have an IQ that is fairly high to do good research work. I don't think that there is any good engineer or scientist that can get along on an IQ of 100, which is the average for human beings. In other words, he has to have an IQ higher than that. Everyone in this room is considerably above that. This, we might say, is a matter of environment; intelligence is a matter of heredity.

Those two I don't think are sufficient. I think there is a third constituent here, a third component which is the one that makes an Einstein or an Isaac Newton. For want of a better word, we will call it motivation. In other words, you have to have some kind of a drive, some kind of a desire to find out the answer, a desire to find out what makes things tick. If you don't have that, you may have all the training and intelligence in the world, you don't have questions and you won't just find answers. This is a hard thing to put your finger on. It is a matter of temperament

probably; that is, a matter of probably early training, early childhood experiences, whether you will motivate in the direction of scientific research. I think that at a superficial level, it is blended use of several things. This is not any attempt at a deep analysis at all, but my feeling is that a good scientist has a great deal of what we can call curiosity. I won't go any deeper into it than that. He wants to know the answers. He's just curious how things tick and he wants to know the answers to questions; and if ^{he} sees things, he wants to raise questions and he wants to know the answers to those.

Then there's the idea of dissatisfaction. By this I don't mean a pessimistic dissatisfaction of the world - we don't like the way things are - I mean a constructive dissatisfaction. The idea could be expressed in the words, "This is OK, but I think things could be done better. I think there is a neater way to do this. I think things could be improved a little." In other words, there is continually a slight irritation when things don't look quite right; and I think that dissatisfaction in present days is a key driving force in good scientists.

And another thing I'd put down here is the pleasure in seeing net results or methods of arriving at results needed, designs of engineers, equipment, and so on. I get a big bang myself out of proving a theorem. If I've been trying to prove a mathematical theorem for a week or so and I finally find the solution, I get a big bang out of it. And I get a big kick out of seeing a clever way of doing some

engineering problem, a clever design for a circuit which uses a very small amount of equipment and gets apparently a great deal of result out of it. I think so far as motivation is concerned, it is maybe a little like Fats Waller said about swing music - "either you got it or you ain't." If you ain't got it, you probably shouldn't be doing research work if you don't want to know that kind of answer. Although people without this kind of motivation might be very successful in other fields, the research man should probably have an extremely strong drive to want to find out the answers, so strong a drive that he doesn't care whether it is 5 o'clock - he is willing to work all night to find out the answers and all weekend if necessary. Well now, this is all well and good, but supposing a person has these three properties to a sufficient extent to be useful, are there any tricks, any gimmicks that he can apply to thinking that will actually aid in creative work, in getting the answers in research work, in general, in finding answers to problems? I think there are, and I think they can be catalogued to a certain extent. You can make quite a list of them and I think they would be very useful if one did that, so I am going to give a few of them which I have thought up or which people have suggested to me. And I think if one consciously applied these to various problems you had to solve, in many cases you'd find solutions quicker than you would normally or in cases where you might not find it at all. I think that good research workers apply these things unconsciously; that is, they do these things automatically and if they were brought forth into the conscious thinking that here's

a situation where I would try this method of approach that would probably get there faster, although I can't document this statement.

The first one that I might speak of is the idea of simplification. Suppose that you are given a problem to solve, I don't care what kind of a problem - a machine to design, or a physical theory to develop, or a mathematical theorem to prove, or something of that kind - probably a very powerful approach to this is to attempt to eliminate everything from the problem except the essentials; that is, cut it down to size. Almost every problem that you come across is befuddled with all kinds of extraneous data of one sort or another; and if you can bring this problem down into the main issues, you can see more clearly what you're trying to do and perhaps find a solution. Now, in so doing, you may have stripped away the problem that you're after. You may have simplified it to a point that it doesn't even resemble the problem that you started with; but very often if you can solve this simple problem, you can add refinements to the solution of this until you get back to the solution of the one you started with.

A very similar device is seeking similar known problems. I think I could illustrate this schematically in this way. You have a problem ^P here and there is a solution ^S which you do not know yet perhaps over here. If you have experience in the field represented, that you are working in, you may perhaps know of a somewhat similar problem, call it P', which has already been solved and

which has a solution, S'. All you need to do - all you may have to do is to find the analogy from P' here to P and the same analogy from S' to S in order to get back to the solution of the given problem. This is the reason why experience in a field is so important that if you are experienced in a field, you will know thousands of problems that have been solved. Your mental matrix will be filled with P's and S's unconnected here and you can find one which is tolerably close to the P that you are trying to solve and go over to the corresponding S' in order to go back to the S you're after. It seems to be much easier to make two small jumps than the one big jump in any kind of mental thinking.

Another approach for a given problem is to try to restate it in just as many different forms as you can. Change the words. Change the viewpoint. Look at it from every possible angle. After you've done that, you can try to look at it from several angles at the same time and perhaps you can get an insight into the real basic issues of the problem, so that you can correlate the important factors and come out with the solution. It's difficult really to do this, but it is important that you do. If you don't, it is very easy to get into ruts of mental thinking. You start with a problem here and you go around a circle here and if you could only get over to this point, perhaps you would see your way clear; but you can't break loose from certain mental blocks which are holding you in certain ways of looking at a problem. That is the reason why very frequently someone who is quite green to a problem will sometimes

come in and look at it and find the solution like that, while you have been laboring for months over it. You've got set into some ruts here of mental thinking and someone else comes in and sees it from a fresh viewpoint.

Another mental gimmick for aid in research work, I think, is the idea of generalization. This is very powerful in mathematical research. The typical mathematical theory developed in the following way to prove a very isolated, special result, particular theorem - someone always will come along and start generalizing it. He will leave it where it was in two dimensions before he will do it in N dimensions; or if it was in some kind of algebra, he will work in a general algebraic field; if it was in the field of real numbers, he will change it to a general algebraic field or something of that sort. This is actually quite easy to do if you only remember to do it. If the minute you've found an answer to something, the next thing to do is to ask yourself if you can generalize this any more - can I make the same, make a broader statement which includes more - there, I think, in terms of engineering, the same thing should be kept in mind. As you see, if somebody comes along with a clever way of doing something, one should ask oneself "Can I apply the same principle in more general ways? Can I use this same clever idea represented here to solve a larger class of problems? Is there any place else that I can use this particular thing?"

Next one I might mention is the idea of structural analysis of a problem. Supposing you have your problem here and a solution

here. You may have too big a jump to take. What you can try to do is to break down that jump into a large number of small jumps. If this were a set of mathematical axioms and this were a theorem or conclusion that you were trying to prove, it might be too much for me to try to prove this thing in one fell swoop. But perhaps I can visualize a number of subsidiary theorems or propositions such that if I could prove those, in turn I would eventually arrive at this solution. In other words, I set up some path through this domain with a set of subsidiary solutions, 1, 2, 3, 4, and so on, and attempt to prove this on the basis of that and then this on the basis of these which I have proved until eventually I arrive at the path S. Many proofs in mathematics have been actually found by extremely roundabout processes. A man starts to prove this theorem and he finds that he wanders all over the map. He starts off and proves a good many results which don't seem to be leading anywhere and then eventually ends up by the back door on the solution of the given problem; and very often when that's done, when you've found your solution, it may be very easy to simplify; that is, to see at one stage that you may have short-cutted across here and you could see that you might have short-cutted across there. The same thing is true in design work. If you can design a way of doing something which is obviously clumsy and cumbersome, uses too much equipment; but after you've really got something you can get a grip on, something you can hang on to, you can start cutting out components and seeing some parts were really superfluous. You really didn't need them in the first place.

Now one other thing I would like to bring out which I run across quite frequently in mathematical work is the idea of inversion of the problem. You are trying to obtain the solution S on the basis of the premises P and then you can't do it. Well, turn the problem over supposing that S were the given proposition, the given axioms, or the given numbers in the problem and what you are trying to obtain is P. Just imagine that that were the case. Then you will find that it is relatively easy to solve the problem in that direction. You find a fairly direct route. If so, it's often possible to invert it in small batches. In other words, you've got a path marked out here - there you got relays you sent this way. You can see how to invert these things in small stages and perhaps three or four only difficult steps in the proof.

Now I think the same thing can happen in design work. Sometimes I have had the experience of designing computing machines of various sorts in which I wanted to compute certain numbers out of certain given quantities. This happened to be a machine that played the game of nim and it turned out that it seemed to be quite difficult. It took quite a number of relays to do this particular calculation although it could be done. But then I got the idea that if I inverted the problem, it would have been very easy to do - if the given and required results had been interchanged; and that idea led to a way of doing it which was far simpler than the first design. The way of doing it was doing it by feedback; that is, you start with the required result and run it back until - run it through its value

until it matches the given input. So the machine itself was worked backward putting range S over the numbers until it had the number that you actually had and, at that point, until it reached the number such that P shows you the correct way. Well, now the solution for this philosophy which is probably very boring to most of you. I'd like now to show you this machine which I brought along and go into one or two of the problems which were connected with the design of that because I think they illustrate some of these things I've been talking about.

In order to see this, you'll have to come up around it; so, I wonder whether you will all come up around the table now.

COVER SHEET FOR TECHNICAL MEMORANDUM

SUBJECT: The Relay Circuit Analyzer - Case 22108

COPIES TO:

CASE FILE

DATE FILE

AREA CENTRAL FILES (4)

- 1 - Patent Dept. (2)
2 - R. Bown
3 - W. H. Doherty
4 - H. H. Abbott
5 - A. O. Adam
6 - A. E. Anderson
7 - E. G. Andrews
8 - M. M. Atalla
9 - H. W. Bode
10 - C. Breen
11 - C. E. Brooks
12 - E. Bruce
13 - A. Burkett
14 - A. J. Busch
15 - R. L. Carmichael
16 - A. B. Clark
17 - C. Clos
18 - R. C. Davis
19 - J. W. Dehn
20 - T. C. Dimond
21 - K. S. Dunlap
22 - F. S. Entz
23 - J. H. Felker
24 - J. G. Ferguson
25 - E. B. Ferrell
26 - G. E. Fessler
27 - W. O. Fleckenstein
28 - J. B. Fisk
29 - G. R. Frost
30 - T. C. Fry
31 - E. N. Gilbert
32 - G. W. Gilman
33 - K. Goldschmidt
34 - R. E. Hersey
35 - B. D. Holbrook
36 - A. W. Horton, Jr.
37 - L. W. Hussey
38 - P. Husta
39 - A. E. Joel, Jr.
40 - M. Karnaugh

FILING SUBJECT
(TO BE ASSIGNED BY AUTHOR)
Switching Theory

~~ABSTRACT~~

- 41 - A. C. Keller
42 - W. Keister
43 - G. V. King
44 - F. A. Korn
45 - W. J. Laggy
46 - C. Y. Lee
47 - E. C. Lee
48 - W. D. Lewis
49 - C. A. Lovell
50 - F. K. Low
51 - A. A. Lundstrom
52 - M. E. Maloney
53 - C. H. McCandless
54 - B. McKim
55 - B. McMillan
56 - B. McWhan
57 - G. H. Mealy

MM-53-1400-9
MM-53-1800-17
DATE March 31, 1953
AUTHOR C. E. Shannon
E. F. Moore

- 58 - J. Meszar
59 - C. G. Miller
60 - D. Mitchell
61 - E. F. Moore
62 - O. J. Murphy
63 - O. Myers
64 - P. B. Myers
65 - N. D. Newby
66 - G. A. Pullis
67 - W. T. Rea
68 - A. E. Ritchie
69 - R. W. Roberts
70 - C. Rosenthal
71 - J. P. Runyon
72 - R. M. Ryder
73 - H. N. Seckler
74 - C. E. Shannon
75 - H. S. Shapiro
76 - F. F. Shipley
77 - F. J. Singer
78 - D. Slepian
79 - L. J. Stacy
80 - R. E. Staehler
81 - E. E. Sumner
82 - F. W. Tatum
83 - J. G. Tryon
84 - S. H. Washburn
85 - E. F. Watson
86 - A. Weaver
87 - W. Whitney
88 - I. G. Wilson
89 - P. L. Wright

(See next page for Abstract)

MM-53-1400-9
MM-53-1800-17
March 31, 1953

ABSTRACT

This memorandum describes a machine (made of relays, selector switches, gas diodes, and germanium diodes) for analyzing several properties of any combinational relay circuit which uses four relays or fewer.

This machine, called the relay circuit analyzer, contains an array of switches on which the specifications that the circuit is expected to satisfy can be indicated, as well as a plugboard on which the relay circuit to be analyzed can be set up.

The analyzer can (1) verify whether the circuit satisfies the specifications, (2) make certain kinds of attempts to reduce the number of contacts used, and also (3) perform rigorous mathematical proofs which give lower bounds for the numbers and types of contacts required to satisfy given specifications.

The Relay Circuit Analyzer - Case 22108

MM-53-1A00-9
MM-53-1800-17
March 31, 1953

MEMORANDUM FOR FILE

1. Introduction

Some operations which assist in the design of relay circuits or other types of switching circuits can be described in very simple form, and machines can be constructed which perform them more quickly and more accurately than a human being can. It seems possible that machines of this type will be useful to those whose work involves the design of such circuits. This is the first of two memoranda describing particular machines of this kind which have been built.

The present machine, called the relay circuit analyzer, is intended for use in connection with the design of two terminal circuits made up of contacts on at most four relays.

The principles upon which this machine are based are not limited to two terminal networks or to four relays, although an enlarged machine would require more time to operate. Each addition of one relay to the circuits considered would approximately double the size of the machine and quadruple the length of time required for its operation.

This type of machine is not applicable to sequential circuits, however, so it will be of use only in connection with parts of the relay circuits which contain contacts, but no relay coils.

2. Operation of the Machine

The machine, as can be seen from Photograph 196492, contains sixteen 3-position switches, which are used to specify the requirements of the circuit. One switch corresponds to each of the $2^4=16$ states in which the four relays can be put. Switch No. 2 in the upper righthand corner, for instance, is labeled $W + X + Y' + Z$, which corresponds to the state of the circuit in which the relays labeled W, X, and Z are operated, and the relay labeled Y is released.

The three positions of this switch correspond to the requirements which can be imposed on the condition of the circuit when the relays are in the corresponding state. Since any single relay contact circuit assumes only one of two values (open or closed), the inclusion of a third value (doesn't matter, don't care, or vacuous, as it has been called by various persons) merits some explanation. If the machine, of which the relay circuit being designed is to be a part, only permits these relays to take on a fraction of the 2^n combinations of which n relays are capable, then (except when considering what the machine will do in case of relay failures) any circuits which agree on the combinations actually assumed will be equivalent in their properties. Since the class of circuits which agree with what is wanted just in the necessary combinations is larger than the class of those which agree in all combinations, the former class can and frequently will contain members using fewer contacts. Hence the switch corresponding to each state is put into the don't care position if the circuit will never assume that state, or if for any other reason the behavior when in that state is immaterial. The sixteen 3-position switches thus permit the user not only to require the circuit under consideration to have exactly some particular hindrance function, but also allow the machine more freedom in the cases where the circuit need not be specified completely.

In order to make a machine of this type to deal with n relays, (this particular machine was made for the case $n = 4$) 2^n such switches would be required, corresponding to the 2^n states n relays can assume. In each of these states the circuit can be either open or closed, so there are 2^{2^n} functionally distinct circuits. But since each switch has 3 positions, there are 3^{2^n} distinct circuit requirements specifiable on the switches, which in the case $n = 4$ amounts to 43,046,721. Thus, the number of problems which the analyzer must deal with is quite large, even in the case of only four relays.

The left half of the front panel of the machine (See Photograph No. 196492) is a plugboard on which the circuit being analyzed can be represented. There are three transfers from each of the four relays, W, X, Y, and Z brought out to jacks on this panel, and two plugs representing the terminals of the network are at the top and bottom. Using these, as well as some patch cords, it is possible to plug up any circuit using at most three transfers on each of the four relays. This number of contacts is sufficient to give a circuit representing any switching function of four variables.

If the specifications for the circuit have been put on the sixteen switches, and if the circuit has been put on the plugboard, the relay circuit analyzer is then ready to operate.

With the main control switch and the evaluate-compare switch both in the evaluate position, pressing the start button will cause the analyzer to evaluate the circuit plugged in, i.e., to indicate in which of the states the circuit is closed by lighting up the corresponding indicator lamps.

Turning the evaluate-compare switch to compare position, the analyzer then checks whether the circuit disagrees with the requirements given on the switches. A disagreement is indicated by lighting the lamp corresponding to the state in question. If a switch is set for closed and the actual circuit is open in that state, or vice versa, a disagreement is indicated, but no disagreement is ever registered when the switch is set in the don't care position, regardless of the circuit condition.

After a circuit has been found which agreed entirely with the requirements, the main control switch is then turned to the short test position and the start button is pressed again. The machine then determines whether any of the contacts in this circuit could have been shorted out, with the circuit still satisfying the requirements. The machine indicates on the lamps beside the contacts which ones have this property.

It may be surprising to the reader than anyone would ever need the assistance of a machine to find a contact which could be shorted out without affecting its circuit. While this is certainly true of simple examples, in more complicated circuits such redundant elements are often far from obvious, particularly if there are some states for which the switches are in the don't care position, since the simplified circuit may be functionally different from the original one, as long as it differs only in the don't care state. It is often quite difficult to see the simplification in these cases.

The analyzer is also helpful in case the circuit being analyzed is a bridge, because of the complications involved in tracing out all paths in the bridge. The circuit shown in Figure 1 is an example of a circuit which was not known to be inefficiently designed until put on the analyzer. It determined in less than two minutes (including the time required to plug the circuit into the plugboard) that one of the contacts shown can be shorted out. How likely would a human being be to solve this same problem in the same length of time?

After the short test has been performed, putting the main control switch in the open test position permits the analyzer to perform another analogous test, this time opening the contacts one at a time.

These two particular types of circuit changes were chosen because they are easy to carry out, and whenever successful, either one reduces the number of contacts required. There are other types of circuit simplification which it might be desirable to have a machine perform, including various rearrangements of the circuit. These would have required more time as well as more equipment to perform, but would probably have caused the machine to be more frequently successful in simplifying the circuit. Using such techniques, it might be possible to build a machine which could design circuits efficiently starting from basic principles, perhaps by starting with a complete Boolean expansion for the desired function and simplifying it step by step. Such a machine would be rather slow (unless it were built to operate at electronic speeds, and perhaps even in this case), and not enough planning has been done to know whether such a machine is practically feasible, but the fact that such a machine is theoretically possible is certainly of interest, whether anyone builds one or not.

Another question of theoretical interest is whether a logical machine could be built which could design an improved version of itself, or perhaps build some machine whose over-all purpose was more complicated than its own. There seems to be no logical contradiction involved in such a machine, although it will require great advances in the general theory of automata before any such project could be confidently undertaken.

To return to the relay circuit analyzer, a final operation which it performs is done with the main control switch in the prove position. Pressing the start button and moving the other 4-position switch successively through the W, X, Y, and Z positions, then certain of the eight lamps W, W', X, X', Y, Y', Z, Z' will light up. The analyzer has carried out a proof as to which kinds of contacts are required to synthesize the function using the method of reduction to functions of one variable, which will be explained in a forthcoming memorandum. The analyzer here ignores whatever circuit has been plugged in the plugboard, and considers only the function specified by the sixteen 3-position switches. If every circuit which satisfies these specifications requires a back contact on the W relay, the W' light will go on, etc.

If, for instance, seven of the eight lights are on, any circuit for the function requires at least seven contacts, and if there is in fact a circuit which uses just seven, the machine has, in effect, given a complete proof that this circuit is minimal. Circuits for which the machine can give such a complete proof are fairly common, although there are also circuits (which can be shown to be minimal by more subtle methods of proof) which this machine could not prove minimal. An example is the circuit of Figure 1. This can be simplified by the analyzer to a circuit of nine contacts, but in the prove position the analyzer merely indicates that at least eight contacts are necessary. It can be shown by other methods that the 9-contact circuit is minimal. But at any rate, the analyzer always gives a mathematically rigorous lower bound for the number of contacts.

3. The Circuit and Operation of the Relay Circuit Analyzer

A complete circuit diagram of the analyzer is shown in Figures 2 and 3. The circuit, as already mentioned, has five modes of operation; 1. evaluating a circuit, 2. comparing a circuit with desired characteristics, 3. examining a circuit for contacts that can be shorted without affecting operation, 4. examining for contacts that can be opened without affecting operation, and 5. proving that certain contacts are necessary in any realization of the function. The method of operation of the circuit will be described in turn for each of these five modes of behavior.

4. Evaluation of a Circuit

In this mode of operation the machine goes through in sequence the sixteen possible states of the relays W, X, Y and Z, that are involved in the circuit and tests in each state whether or not the circuit is closed. If it is closed, the corresponding panel light is lit. In this process only the right-hand part of the circuit in Figure 2 is involved and switches S18 and S19 are both in the evaluate position. The selector switch S17 goes through one complete revolution to make this test. During this revolution the four relays W, X, Y, and Z proceed sequentially through their sixteen states. This sequence is produced by the first two wipers and decks of the selector switch S17. At the first position (0000) all four relays are unoperated. At the second step (0001), ground on the second wiper operates relay Z, which locks in on its own front contact. The circuit is then set to test the situation where W, X and Y are unoperated and Z is operated. At the third step relay Y is operated and locks in on

its own front contact. At the fourth step Z is short-circuited by the wiper of the first deck. This releases Z and produces the state 0010. Proceeding in this manner it will be seen that the four relays W, X, Y and Z go through the sixteen states indicated. The circuit which is being tested may be thought of as being connected between plugs P1 and P2 at the upper left of the diagram. This network consists of contacts on the four relays W, X, Y and Z. Actually some other contacts are involved in the network between P1 and P2 (contacts on the H relays) but in the present mode of operation these H relays do not operate and do not affect the hindrance from P1 to P2. For a given state of the relays W, X, Y and Z the plugs P1 and P2 will be connected together if, and only if, the circuit being tested is closed for that state of the relays. The relay G will, therefore, operate if, and only if, the circuit is closed in the state in question. If it is closed, a ground will be applied to the third wiper of the selector switch S17 and this will fire the corresponding neon lamp. If it is not closed +84 volts will be applied to the lamp extinguishing it (if it is already fired). The voltage across the lamp circuit, 84-24 or about 60 volts, lies between the fire and sustain voltages for the neon lamps. Consequently, if they are fired they will remain fired, if extinguished they will remain out. Thus the lamps remain in the state produced by the evaluation of the circuit even after the wiper has left the point in question.

The movement of the stepping switch is produced by a three-stage buzzer circuit consisting of relays U, V and P. In the buzzing condition the parallel S' and T' combination in series with U will be closed. The operation of U energizes V through the front U contact in series with the V coil. The operation of V then operates P in a similar manner. The operation of P releases U through the P' contact. This releases V which releases P, etc.

At the start of an evaluation, switch S18 will be in the evaluate position, switch S19 in the evaluate position, selector switch S17 at position 22 (and relay S, therefore, operated) and selector switch S16 at position 21 (with relay T, therefore, operated). When the starting push button S20 is pressed magnet M1 of stepping switch 1 is energized. When S20 is released M1 releases and the stepping switch moves to position one. This releases relay S and the three-stage buzzer U, V, P starts operating. At each cycle of this buzzer the coil of selector switch S17 is energized and released by a make contact on the P relay. This sequences the relays W, X, Y and Z through their sixteen states, as already described, and indicates on the neon lamps the states for which the circuit being tested is closed. When the wipers reach level 22 relay S operates, stopping the buzzer and ending the test.

5. The Comparison Mode of Operation

In this mode of operation the circuit set up on the plugboard is to be compared with the settings of the sixteen three-position switches. If in any state the circuit disagrees with the switch setting the corresponding neon lamp will light up. For this test switch S18 is set in the evaluate position and switch S19 in the compare position. When the starting push button S20 is pressed, the buzzing circuit U, V, P starts as before, cycling the selector switch S17 through one complete revolution. The four relays, as before, go through their sixteen possible states and the relay G, as before, operates or not, depending on whether the circuit being tested is closed or not. The lamps, however, are no longer controlled directly by the relay G, but instead by contacts on the relay A. The relay A is connected to operate, if, and only if, the circuit condition of the network being tested (open or closed) disagrees with the setting of the corresponding three-position switch. This result is obtained by having one end of the coil of relay A connected (via the fourth wiper of selector switch S17) to +24 volts, nothing (i.e. floating) or minus, according to the desired behavior of the circuit in the state in question is open, "don't care", or closed (as represented by the setting of the three-position switch). The other end of the relay A is connected to +24 volts or minus, according as the actual circuit under test is open or closed (this being carried out by a transfer on the G relay). The relay A will operate only if the two ends of the coil receive different polarities, and this will occur only if the switch setting differs from the state of the network under test as indicated by the state of the relay G. If such a disagreement occurs the corresponding lamp is fired by a ground coming in the third wiper of selector switch S17.

The starting and stopping are carried out by the same means as used in the evaluate mode.

6. The Short Test

In testing for contacts in the circuit that can be shorted, the sequencing is somewhat more involved. Roughly speaking, the various contacts used in the circuit are short-circuited one-by-one, and for each contact the circuit goes through a sequence similar to the comparing mode of behavior just described (comparing the circuit when this contact is shorted with the desired characteristics set up on the three-position switches). If any disagreement is found, the neon lamp associated with the contact in question is fired, indicating that this contact is necessary in the circuit and cannot

he shorted. Actually, the sequence is a bit more complicated since to save time and equipment the tests on the make and break parts of a transfer in the circuit being tested are interleaved.

To carry out the short test switch S18 is put in the short position (the position of S19 is irrelevant). The selector switches S16 and S17 start in positions 21 and 22 respectively, so that relays 3 and T are both operated. When the starting button S20 is pressed, the magnets of both S16 and S17 are energized and when S20 is released they step ahead one step releasing both S and T and allowing the buzzer circuit to start. The first step of selector switch S16 causes E to operate. This removes the voltage from the indicating lamps L16 to L39 (removing any indication on these lamps from previous runs). Stepper 1 then proceeds through a complete revolution. At step 17 the second wiper applies a voltage to the coil of S16, pulsing S16 ahead one notch. This releases E, and reapplies voltage to the indicating lamps L16 to L39. The wipers of selector switch S16 are now connected to position 1 (the top row) of this selector. The sixth wiper operates relay H1 which disconnects the first W transfer from the circuit being tested. The three points in the circuit being tested that were previously connected to this transfer (on the W relay) are brought down to points P3, P5 and P7, P5 coming through the third wiper. The free ends of the W transfer, that are now disconnected from the circuit being tested are brought down via wipers 2 and 4. To test whether either part of this transfer can be shorted, the selector switch S17 goes through a complete cycle, putting the relays W, X, Y and Z in each possible state as in previous modes of operation. In each state, the first test is to short P3 to P5, which in effect shorts the nodes of the circuit normally connected to the W part of the contact, and the circuit state is compared with the desired specification on the three-position switch. A disagreement operates relay A which, by way of wiper 1, fires the lamp corresponding to the W contact. This shorting of the nodes occurs in the buzzer cycle during the period when the relay U is operated. The A contact is connected to the corresponding lamp through contact V and P' in series. This gives relay A time to operate (or release from a previous operation) before its reading is applied to the lamp, and also disconnects the lamp before the state of A is changed by the next operation.

The second test in the same buzzing cycle is to short the break contact of the transfer. This occurs when U releases, connecting P3 to P4 and P5 to P7. The W make is then connected as usual in the circuit being tested (via the

H1 make, U', and wiper 2) and the nodes previously connected to the back W' contact are shorted via the 3rd wiper of selector switch S16. In this part of the buzzing cycle the disagreement relay is connected via P and V' contacts (for timing margins similar to P' and V before) and the 5th wiper, to the lamp corresponding to the W', or break contact. This lamp will fire, as before, if a disagreement occurs indicating that the contact is necessary.

After selector switch S17 has run through all states (rows 1 to 16) it applies ground through wiper 2 to the magnet of selector switch S16, advancing it one step. The machine now applies the shorting test to the X and X' contacts connected to the second row of selector switch S16. Proceeding in this manner it tests all the contacts. On reaching row 13, the 6th wiper of selector S16 applies ground to its own coil through its own back contact. This causes it to step rapidly through the remaining positions until it reaches row 21 where it operates relay T. The first selector switch is meanwhile still being pulsed by the buzzer circuit. After T operates, the first time S17 reaches row 22, relay S operates and the buzzer stops. This completes the test.

If it is desired to hurry the machine through the latter part of a test (for example if only a few of the available contacts are being used and these are near the top) the reset button S21 can be pressed. This causes S16 to run rapidly to the stop position (row 21).

7. The Open Test

The test for opening contacts proceeds exactly as the short test just described, except that having switch S18 in the open position opens wiper 3 of S16. This opens the short that was applied in the previous test to the nodes normally connected to the contact being tested. The relay therefore indicates the behavior of the circuits when the different contacts are opened.

8. The "Prove" Mode of Operation

When switch S18 is set in the "prove" position the machine indicates, by lighting some of the lamps L40 to L47, that certain contacts are necessary in any circuit which realizes the switching function set up on the sixteen three-position switches. This indication is obtained by moving switch S22 through its four possible positions. In the W position the machine tests whether W and/or W' contacts are necessary and if so, lights the corresponding lamps etc.

The method of operation is based on the following result in switching theory (stated for simplicity for the case of four variables). At least one W (make) contact is necessary in any realization of a given switching function if there are one or more states of the other relays (X, Y, and Z) such that when the X, Y and Z relays are in such a state, changing the W relay from unoperated to operated changes the function from open to closed. At least one W' (break) contact is necessary if there exists a state of the X, Y and Z relays such that when they are in this state, operating the W relay changes the circuit from closed to open. These are both obvious, since the only way by which operating the W relay alone could close a previously open circuit is by establishing an operating path through a make contact on the W relay, and similarly for the condition with a break contact.

The condition that a W contact is necessary can also be thought of geometrically in the following way. The sixteen states of the four relays can be thought of as the vertices of a four-dimensional cube. This cube consists of two three-dimensional subcubes, the first being the eight states of the X, Y, Z relays with W not operated, and the second, the eight states of the X, Y, Z relays with W operated. If there is any point in the "W unoperated" cube in which the circuit is open (closed) while being closed (open) in the corresponding point of the "W operated" cube, at least one W (W') contact is necessary.

The "Prove" part of the circuit can best be understood in terms of this geometrical picture. A two-terminal network with terminals a and b is set up in the machine, corresponding to this cube. Every vertex of the cube for which the circuit should be closed is connected to terminal a; all vertices for which the circuit should be open are connected to terminal b ("don't care" vertices are left floating). When testing for the necessity of W or W' contacts, eight diodes are connected between corresponding points of the three-dimensional subcubes mentioned above. These point from the "W unoperated" subcube to the "W operated" subcube. Current will pass from terminal a to terminal b if and only if a W contact is necessary. This is true since this conduction can take place only by entering the cube at a closed state (these being the only ones connected to terminal a), passing through a diode in the conducting direction (this requires that the closed state be in the "W unoperated" cube) and leaving the cube to terminal b at an open state. Thus the conditions for conduction from a to b are identical with the conditions for necessity of a W contact. In a similar manner, it may be seen that the network will conduct from b to a if and only if a W' contact is necessary.

In operation, the circuit is alternately tested for conduction in the two directions. The alternation is obtained by operation of the four-stage buzzer previously described. When P is operated, the circuit is tested for conduction from A to B. If this condition occurs, it fires the corresponding neon lamp (for the W, X, Y or Z make contact). When P is released, voltage is applied to the AB network in the reverse direction and if conduction occurs, it fires the corresponding neon lamp (for the W', X', Y' or Z' break contact). These lamps remain fired until released either by turning off the main power or flipping the "evaluate-compare" switch S19 from one position to the other.

Although it has been explained that the circuit for doing these tests is laid out in the shape of a four-dimensional cube, the circuit diagram of Figure 3 is not drawn by the use of a direct projection of such a cube, but is laid out in a plane by a method due to W. Keister (The Design of Switching Circuits, D. Van Nostrand, 1951, p. 174), which simplifies its appearance.

It can easily be verified that by putting switch S22 in any one of its four positions the circuit in Figure 3 reduces to a 4-dimensional cube with 8 diodes joining its two halves. However the manner in which these 4 sets of 8 diodes each were combined to give a total of only 14, while at the same time using only 8 decks of the switch S22, may be of interest. It can be applied to give similar economies in the design of analogous circuits for cubes of any dimension. This method depends on some concepts due to R. W. Hamming (Bell System Technical Journal, 29, pp.147-160, April, 1950). It is possible to divide the vertices of an n-cube into two mutually exclusive and collectively exhaustive classes, called parity classes, depending on whether the number of coordinates having the value 1 is even or odd. If a point belongs to one parity class, all of the points which have distance 1 from it (and hence differ in only one coordinate from it) are in the opposite parity class. This means that every edge of the cube connects vertices of opposite parity classes. Since in every position of S22 the diodes are connected along edges of the cube, it means that it is necessary to be able to connect diodes only between points of opposite parity classes.

Thus the diodes are all connected to the points of one parity class, and the decks of switch S22 are connected to the points of the other class. If one diode pointing toward and one pointing away from each point of the even parity class is provided, then the switch contacts can connect each point of the other parity class to the other end of the proper one of these two diodes. In the actual circuit not quite this many diodes are used, since the points 0000 and 1111 require only one of the two diodes.

9. Notes and Comments

The small size and portability of this machine depend on the fact that a mixture of relay and electronic circuit elements were used. The gas diodes are particularly suited for use where a small memory element having an associated visual display is required, and the relays and selector switches are particularly suited for use where the ability to sequence and interconnect using only a small weight and space is required. In all, the relay circuit analyzer uses only 24 relays, 2 selector switches, 48 miniature gas diodes, and 14 germanium diodes as its logical elements.

It may be of interest to those familiar with general purpose digital computers to compare this method of solution of this problem on such a small, special-purpose machine with the more conventional method of coding it for solution on a high-speed general-purpose computer. One basic way in which the two methods differ is in the directness with which the circuits being analyzed are represented. On a general-purpose computer it would be necessary to have a symbolic description of the circuit, probably in the form of a numerical code describing the interconnections of the circuit diagram, and representing the types of contacts that occur in the various parts of the circuit by means of a list of numbers in successive memory locations of the computer. On the other hand, the relay circuit analyzer represents the circuit in a more direct and natural manner, by actually having a copy of it plugged up on the front panel.

This difference in the directness of representation has two effects. First, it would be somewhat harder to use the general-purpose computer, because the steps of translating the circuit diagram into the coded description and of typing it onto the input medium of the computer would be more complicated and lengthy than the step of plugging up a circuit directly. The second effect is in the relative number of logical operations (and hence, indirectly, the time) required by the two kinds of machines. To carry out the fundamental step in this procedure of determining whether the given circuit (or some modification of it obtained by opening or shorting a contact) is open or closed for some particular state of the relays requires only a single relay operate time for the relay circuit analyzer. However, the carrying out of this fundamental step on a general-purpose digital computer would require going through several kinds of subroutines many times. There would be several ways of coding the problem, but in a typical one of them the computer would first go through a subroutine to determine whether a given contact were open or closed, repeating this once for each contact in the circuit,

and then would go through another subroutine once for each node of the network. Altogether this would probably involve the execution of several hundred orders on the computer, although by sufficiently ingenious coding this might be cut down to perhaps 100. Since each order of a computer takes perhaps 100 times the duration of a single logical operation (i.e., a pulse time, if the computer is clock-driven), it turns out that what takes 1 operation time on one machine takes perhaps 10,000 on another.

Since 10,000 is approximately the ratio between the speed of a relay and of a vacuum tube in performing logical operations, this gain of about 10,000 from the directness of the representation permits this relay machine to be as fast as a general-purpose electronic computer.

This great disparity between the speeds of a general-purpose and of a special-purpose computer is not typical of all kinds of problems, since a typical problem in numerical analysis might only permit of a speed-up by a factor of 10 on a special-purpose machine (since multiplications and divisions required in the problem use up perhaps a tenth of the time of the problem). However, it seems to be typical of combinatorial problems that a tremendous gain in speed is possible by the use of special rather than general-purpose digital computers. This means that the general-purpose machines are not really general in purpose, but are specialized in such a direction as to favor problems in analysis. It is certainly true that the so-called general purpose machines are logically capable of solving such combinatorial problems, but their efficiency in such use is definitely very low. The problems involved in the design of a general-purpose machine suitable for a wide variety of combinatorial problems seem to be quite difficult, although certainly of great theoretical interest.

10. Conclusion

An interesting feature of the relay circuit analyzer is its ability to deal directly with logical circuits in terms of 3-valued logic. There would be considerable interest in techniques permitting easy manipulation on paper with such a logic, because of its direct application to the design of economical switching circuits. Even though such techniques have not yet been developed, machines such as this can be of value in connection with 3-valued problems.

Whether or not this particular kind of machine ever proves to be useful in the design of practical relay circuits, the possibility of making machines which can assist in logical design procedures promises to be of value to everyone associated with the design of switching circuits. Just as the slide rule and present-day types of digital computers can help perform part of the routine work associated with the design of linear electrical networks, machines such as this may someday lighten much of the routine work associated with the design of logical circuits.

C. E. SHANNON

E. F. MOORE

Attached:
Photograph No. 196492
Figures 1, 2 and 3

FIGURE 1

THE RELAY CIRCUIT ANALYZER WAS ABLE TO SIMPLIFY
THIS CIRCUIT, REMOVING ONE CONTACT, IN LESS THAN
TWO MINUTES TOTAL TIME. CAN YOU DO AS WELL?

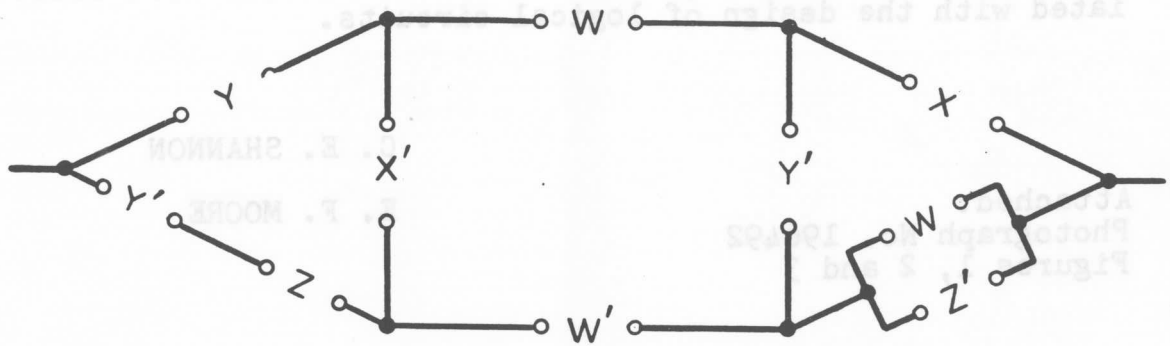
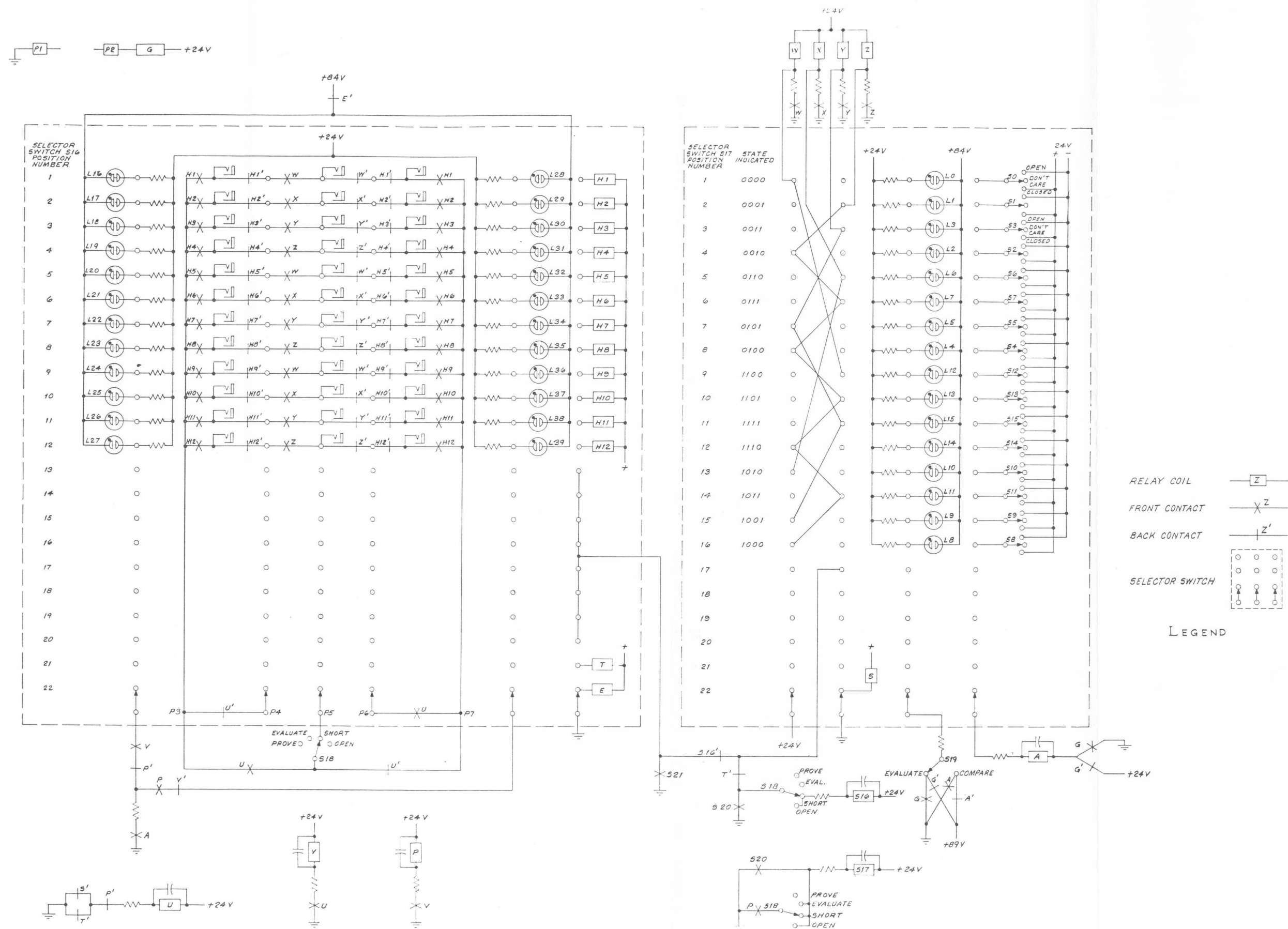
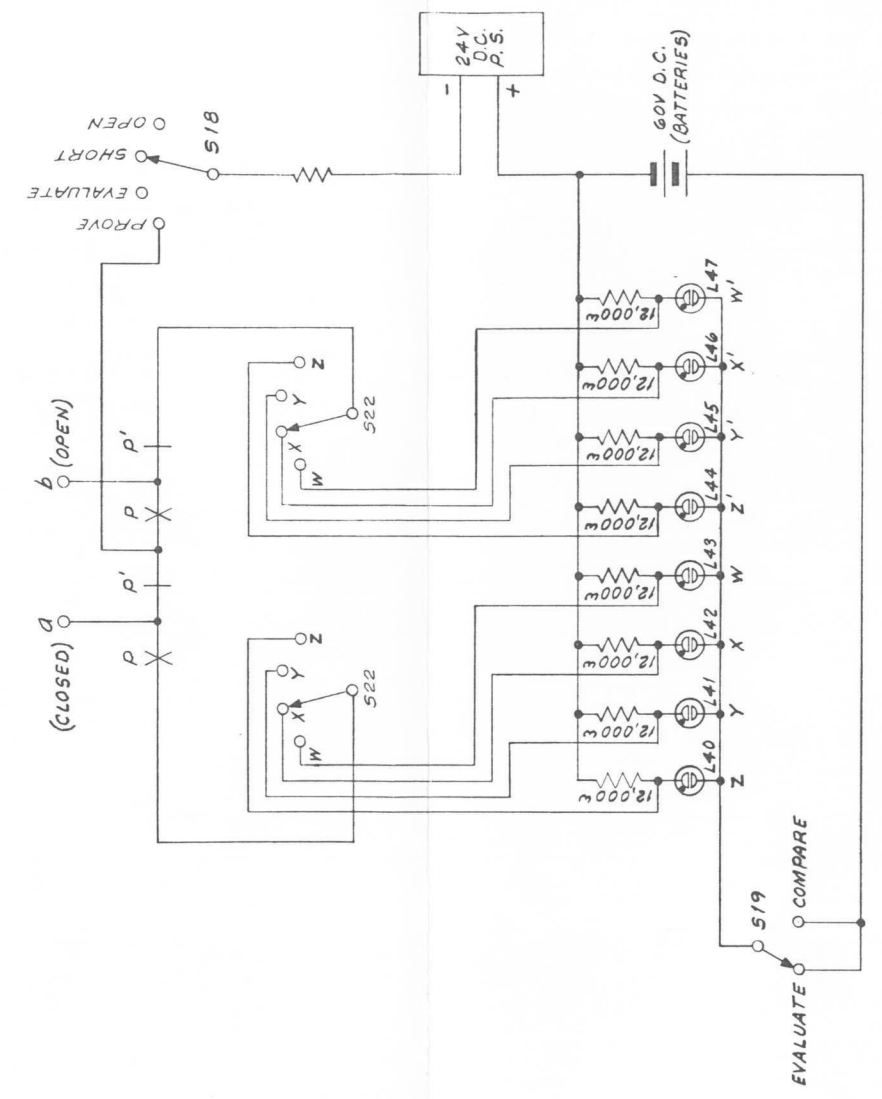
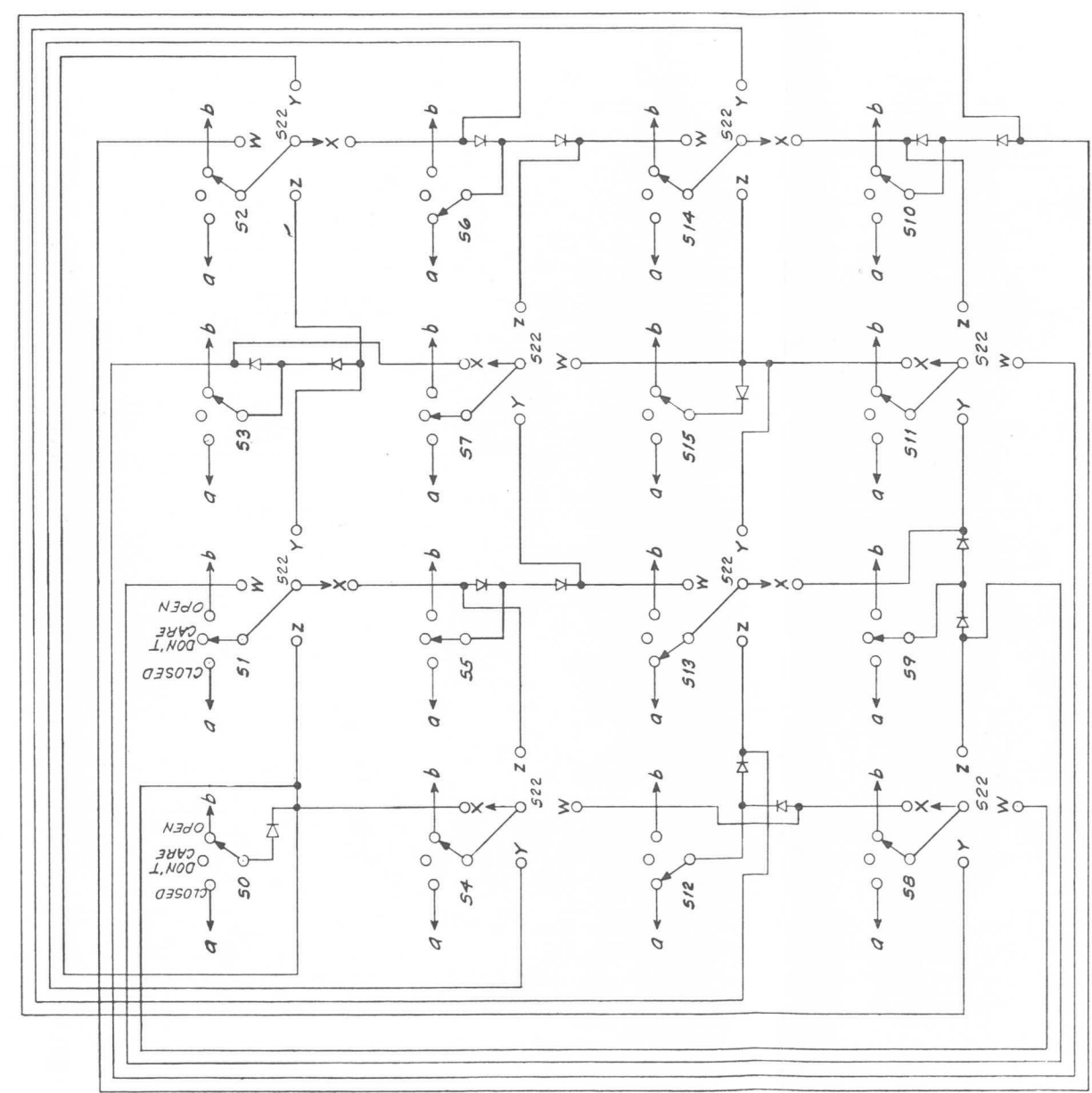
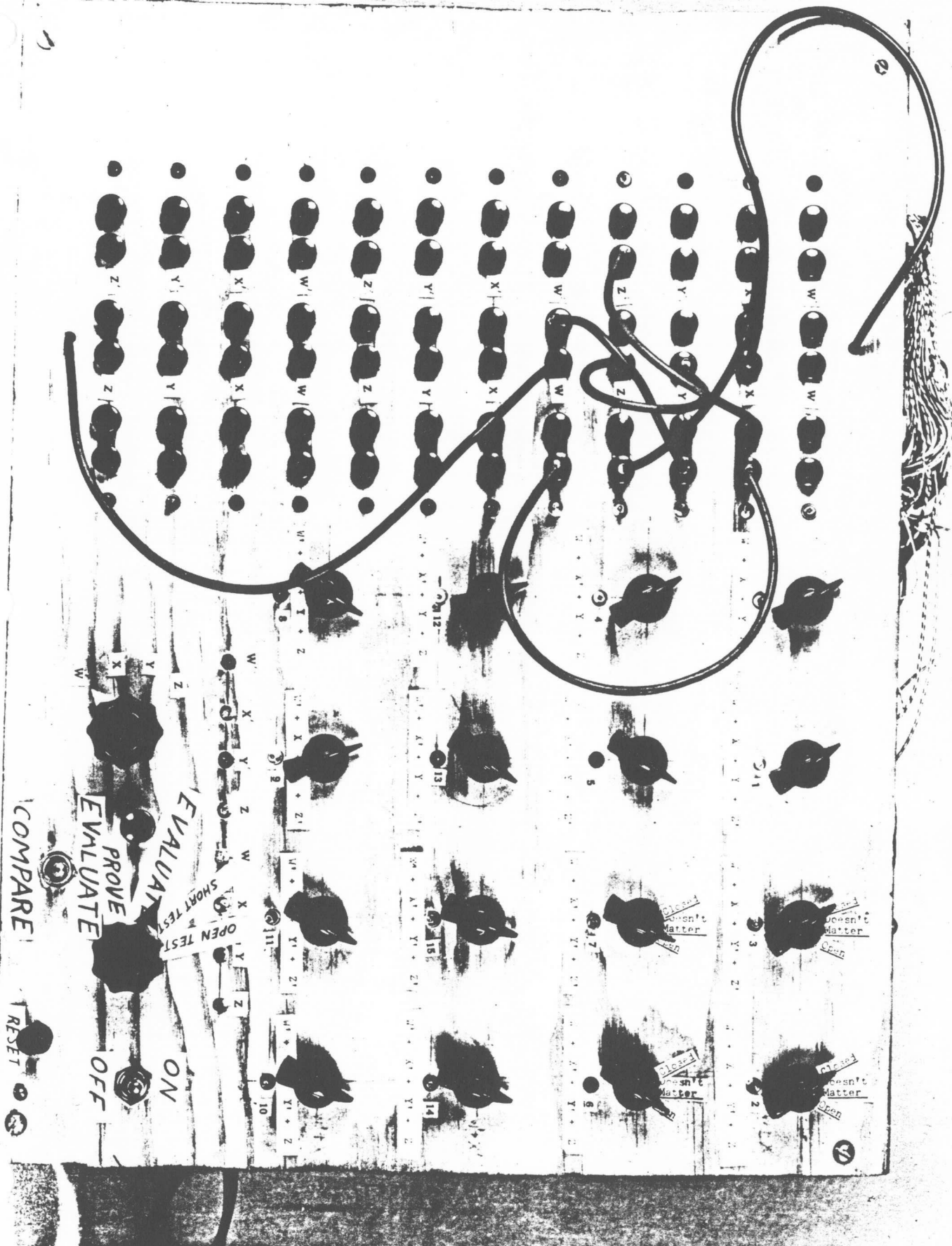


FIGURE I

THE RELAY CIRCUIT ANALYZER WAS ABLE TO SIMPLIFY THIS CIRCUIT, REMOVING ONE CONTACT, IN LESS THAN TWO MINUTES TOTAL TIME. CAN YOU DO AS WELL?







THROBAC - CIRCUIT OPERATION

The central part of the Throbac circuit is a relay accumulator which can count up to eighty in a modified Roman numeral system. The accumulator is arranged so that it is possible to add or subtract I, V, X or L to the contents of the accumulator. It consists of seven stages of W-Z circuits. The first three stages W₁-Z₁, W₂-Z₂ and W₄-Z₄ accumulate "I's". These stages are arranged to count up to four and recycle to zero at the fifth I. Thus, within these stages either zero, one, two, three or four "I's" will be registered. The number of "I's" appears in binary form in the three stages of W-Z.

The next W-Z combination W_V-Z_V accumulates "V's", either zero or one V being registered here. The final three stages W_X₁-Z_X₁, W_X₂-Z_X₂ and W_X₄-Z_X₄ accumulate "X's" from zero up to seven.

If the relay F is operated, the accumulator is arranged to add; if F is released, to subtract. Supposing F operated, closing P_I adds I to the contents of the accumulator. Closing P_V adds V, P_X adds X and P_L adds L. This may be verified by tracing out the circuit paths into the W-Z circuits in the various cases. For example, if the accumulator has zero in it, all W's and Z's are released, and when P_I is closed a ground passes through a chain of contacts P_I-F-Z₄-F to pulse the W₁-Z₁ pair, and this is the only W-Z pair to receive a ground. If, instead, P_L had been pulsed, the W_X₁-Z_X₁ pair and the W_X₄-Z_X₄ pair would both receive ground, thus registering L (XXXX + X). A study of the circuit

will show that in all cases it adds or subtracts (according to the state of F) I, V, X or L when P_I , P_V , P_X or P_L is operated.

At the bottom of this circuit a connection leads out to control the C relay. This connection will be seen to carry a ground when a number is added to the accumulator which causes it to overrun its limit either by addition, giving a number greater than seventy-nine, or, by subtraction, a number less than zero. In these cases the carrying to or borrowing from what would be the next column goes out on the lead in question to control the C relay. This relay, to be described later, indicates the end of a division.

The number registered in the accumulator is displayed on the panel by means of a series of thirteen lights. These lights are controlled by contact networks on the W-Z relays of the accumulator. The contact networks translate from the modified Roman numeral notation to the standard one. The part of the number which is a multiple of ten appears in the three left columns of lights, L_7 or X_7 , L_6 or X_6 , L_5 or X_5 . The part of the number registered which is less than ten appears in the four right columns of lights.

As an example, suppose the number registered is LXIV (64). In the accumulator the W-Z pairs W_4-Z_4 (IIII), W_X-Z_X and W_X-Z_X (XXXXX) will be operated and other W-Z pairs released. In the accumulator light circuit it will be found that lights L_6 , X_5 , I_4 and V_3 will receive a ground and be illuminated, displaying the number LXIV.

The sequencing for adding or subtracting a number entered in the keyboard into the accumulator is carried out chiefly by stepping switch A. For such an addition or subtraction, this stepper sweeps across the keyboard, starting from the right-hand column and sequentially adding or subtracting the numbers registered in each column. The addition sequence is started by pressing the ADD button which causes P to operate and lock in through a back contact on K. The operation of P causes the buzzer relay D to start operating and releasing at about ten cycles per second. When D closes it pulses the stepping coil of stepper A, moving it ahead one notch. The release of D puts a ground on the wipers of the stepper and, therefore, on the first vertical connection through the keyboard switches. Let us suppose that the number XLVI is entered in the keyboard in the four right-hand columns. I is then registered in the right most column and the ground from the stepper passes through this I push button to operate the P_I relay. The F relay has been operated by P and therefore I is added to the previous contents of the accumulator. On the next cycle of the buzzer, the stepper moves to the next column and operates the P_V relay which adds V into the accumulator. P_V also causes R to operate and lock in through K' . The purpose of this is to cause any further I's to be subtracted rather than added. On the next cycle of the buzzer, ground is applied to the third vertical of the keyboard and, because of the L entered there, operates the P_L relay. This adds L to the accumulator and also operates the S relay, which also

locks in through K'. The operation of S signifies that an I has occurred and consequently any X's or V's now encountered on the keyboard must be subtracted. On the next cycle of the buzzer, the fourth vertical receives ground and because of the X in this column, P_X operates. Since S is closed, the relay M also operates, releasing F and making the accumulator subtract instead of add. The timing of these relays is adjusted so that F releases before the P_X pulse could add into the accumulator. X is therefore subtracted. On the next three cycles of the buzzer, no further numbers are encountered and the accumulator does not change. On the eighth cycle, the wipers pass a ground to the K relay which locks in momentarily, and also to the reset coil of the stepper. The operation of K releases relays P, R and S and also disconnects the buzzer and the wipers. The reset coil allows the wipers to return to their normal position and since they have been disconnected by K they have no effect as they pass over the keyboard columns. When the wipers reach their normal position they open the off-normal switch of the stepper. This releases K and the addition operation is complete.

The process of subtraction is essentially the same. Pressing the subtract button causes M to operate and lock up, which starts the buzzer and the stepping operations. In this case, however, F is normally released, so that numbers encountered in the keyboard are normally subtracted. However, when a smaller number is encountered after a larger one the relay F will operate, causing it to be added.

Multiplication is obtained by successive addition. If the MV button is pressed, the machine adds the contents of the keyboard into the accumulator V times, if the MX button is pressed X times. This counting is controlled by stepper B. If the MI button is pressed, the keyboard contents are added or subtracted depending on whether the MV or MX buttons have been previously operated.

Suppose VIII is to be multiplied by IV. VIII is entered in the keyboard and first the MV and then the MI push buttons pressed. When the MV button is pressed, relay MV operates and locks in through Q'. The relay T also operates, locking in through the Clear Upper key. The relay T signifies that I's occurring later in the multiplier must be interpreted as negative. The operation of MV causes the P relay to operate and start an addition operation. When stepper A reaches the eighth point, K operates causing the stepping coil of stepper B to receive a ground (through the MV make). When stepper A resets to normal, P again operates, again adding the keyboard contents into the accumulator and advancing stepper B at the end of the addition. This process continues until stepper B reaches its fifth point. There the ground on the wipers operates relay Q which releases MV and stops the series of additions. Q locks in and applies ground to the reset coil of stepper B, returning it to normal. When it reaches normal, the off-normal contacts are opened and Q is released.

Next the MI button is pressed. Since T is in (due to the previous operation of MV), this causes M to operate and the machine subtracts the keyboard contents from the accumulator. This completes the multiplication. The MX button produces a sequence similar to the MV button, except that stepper B must go to the tenth point instead of the fifth to operate Q and stop the series of additions.

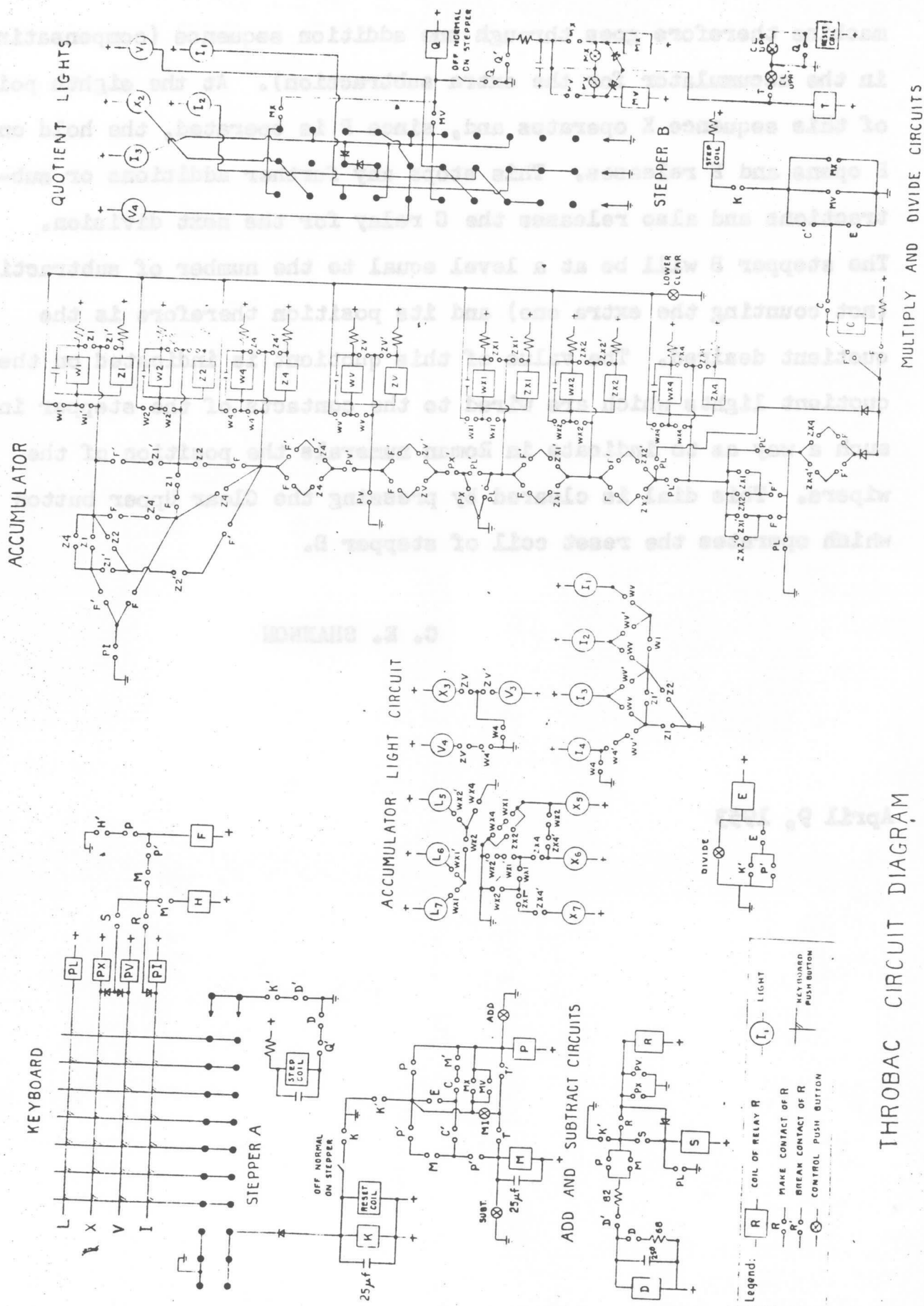
If another multiplication is to be performed, the Clear Upper button should be pressed. This releases T and resets stepper B to normal if for some reason it is not already there.

Division is performed by successive subtraction. The dividend is entered in the accumulator and the divisor in the keyboard. When the divide button is pressed, relay E operates and locks in through P' or K'. C is normally out and E, therefore, causes M to operate and lock in, starting a subtraction. If, during this subtraction, the accumulator does not run through zero, C will not operate and another subtraction will occur since M will again operate as soon as K releases. At each subtraction of this sort the operation of K at the end of the subtraction energizes the stepping coil of stepper B advancing it one step. Eventually in this subtraction process the contents of the accumulator will go negative. This causes C to operate and indicates that one too many subtractions have been performed. The last subtraction is not counted on stepper B since its operating path passes through C'. The operation of C causes the next operation to be an addition, since the next ground when K releases is placed on P rather than M. The

machine therefore goes through one addition sequence (compensating in the accumulator for the extra subtraction). At the eighth point of this sequence K operates and, since P is operated, the hold on E opens and E releases. This stops any further additions or subtractions and also releases the C relay for the next division. The stepper B will be at a level equal to the number of subtractions (not counting the extra one) and its position therefore is the quotient desired. The value of this quotient is indicated on the quotient lights which are wired to the contacts of the stepper in such a way as to indicate in Roman numerals the position of the wipers. This dial is cleared by pressing the Clear Upper button which operates the reset coil of stepper B.

C. E. SHANNON

April 9, 1953



TOWER OF HANOI

C. E. Shannon

The Tower of Hanoi machine automatically solves a well-known puzzle constructed as follows. There are three pegs standing upright in a horizontal plate. On the first peg are a number of disks of graduated sizes. The problem is to move all these disks to the third peg subject to the rules that (1) only one disk can be moved at a time, and (2) a disk can never be placed on top of a smaller disk.

This puzzle has been treated in the literature. It can be readily proved by induction that with n disks, $2^n - 1$ moves are necessary. For suppose this formula is true up to $n-1$. With n disks, in order to move the largest one to the third peg it is necessary that all the other disks be on the second peg in proper order. This, by assumption, requires $2^{n-1} - 1$ moves. Moving the largest disk requires one more and moving the $n-1$ disks from the second to the third peg, again by the inductive hypothesis, requires $2^{n-1} - 1$ moves. Consequently the entire operation requires $2^n - 1$ moves. Since the formula is true for $n = 1$, it holds in general. The argument also shows how to build up a solution for any n from $n-1$, and hence, eventually, from the $n = 1$ case.

For $n = 6$ (the case handled by the machine) the solution is given by the following table.

000000	000000	100000	211111
000001	000001	100001	211112
000010	000021	100010	211102
000011	000022	100011	211100
000100	000122	100100	211200
000101	000120	100101	211201
000110	000110	100110	211221
000111	000111	100111	211222
001000	002111	101000	210222
001001	002112	101001	210220
001010	002102	101010	210210
001011	002100	101011	210211
001100	002200	101100	210011
001101	002201	101101	210012
001110	002221	101110	210002
001111	002222	101111	210000
010000	012222	110000	220000
010001	012220	110001	220001
010010	012210	110010	220021
010011	012211	110011	220022
010100	012011	110100	220122
010101	012012	110101	220120
010110	012002	110110	220110
010111	012000	110111	220111
011000	011000	111000	222111
011001	011001	111001	222112
011010	011021	111010	222102
011011	011022	111011	222100
011100	011122	111100	222200
011101	011120	111101	222201
011110	011110	111110	222221
011111	011111	111111	222222

The first column gives the binary numbers from 0 to 63. The second column describes the positions of the disks. For example, 000000 means that all disks are on peg 0. The fifth entry 000122 means that the three largest disks are on peg 0, the next smaller disk on peg 1, and the two smallest disks on peg 2. The numbers in the second column are related in a peculiar manner to the binary numbers in the first column and can be calculated from them. The process can best be described by an example. Take, for instance, the binary number 010110. The following calculation is

performed.

	+	-	+	-	+	-
	0	1	0	1	1	0
	0	2	2	1	2	2
	0	1	2	0	0	2

The columns here alternate + and -. The second row 022122 is obtained by summing the first row horizontally mod 3 with + or - sign depending on the column. Thus $0=0$, $2=0-1$, $2=0-1+0$, $1=0-1+0-1$, $2=0-1+0-1+1$ and $2=0-1+0-1+1-0$ (all mod 3). The third row is obtained from the second by alternately adding and subtracting the first row from it. This row is the corresponding position of the disks in the solution of the puzzle. It can be shown that this relation holds in general.

The Tower of Hanoi relay circuit is based on this curious relation. The machine basically consists of a binary counter (six stages of W-Z counters) which counts from 0 to 63. Contacts on these relays are connected in a network which controls a set of eighteen lights. There are three lights for each of the six disks, one on each of the three pegs. At a given time, one of these three will be on, indicating the position of the corresponding disk. As the counter proceeds through its count, the lights are switched to indicate the process of the solution.

The circuit of the machine is shown in Fig. 1. The right hand network controls the lights. It will be seen that this consists of a symmetric function lattice in which the stages alternately add and subtract mod 3. The ground coming in at the bottom of this circuit will appear in columns 0', 1', 2' according to the first number computed in the above calculation (i.e. 0'2'2'1'2'2' in the example given). The further calculation (012002 in the example) is carried out by the single stage mod 3 circuits attached to

the basic mod 3 lattice.

It is interesting in this circuit that when one of the larger disks is moved the lamps corresponding to smaller disks receive their operating current through a path which is switched. The counting process, however, is so rapid that they appear to be continuously illuminated.

The control circuit at the left of the figure contains a three-position key switch. In the center position, the machine stops. In the top position, it causes the buzzer B to operate the counter and therefore proceed through the solution at about two steps per second. When the count reaches sixty-three, the buzzer stops. If the key switch is depressed to the lower position (non-locking), the counter is advanced one count. By moving the switch between the center and the lower positions the solution can be observed step by step.

Original was dated April 20, 1953

Mathmanship or How to Give an Explicit Solution Without Actually Solving the Problem

After reading several weighty papers giving formulas which assume only prime values, I felt moved to develop a few further results of the same type.

Theorem 1. There exists a unique real positive number $\lambda < 1$ such that

$$a_n = [2^n \lambda] - 2[2^{n-1} \lambda]$$

$$= \begin{cases} 0 & \text{if } n \text{ is composite} \\ 1 & \text{if } n \text{ is prime} \end{cases}$$

Here $[x]$ means, as usual, the largest integer in x .

The value of λ is .413

Theorem 2. There exists a unique real positive number $\mu < \frac{1}{4}$ such that the n^{th} prime p_n is given by

$$p_n = [2^{2^{n+1}} \mu] - 2^{2^n} [2^{2^n} \mu]$$

Note the improvement over previous results - this formula gives all the primes, not just some of them. For analysts who find the bracket symbol a little suspect, we have the following:

Theorem 3. There exists a real number λ such that $\sin 2^n \lambda$ is positive or negative according as n is prime or composite.

Theorem 4. There exists a real number δ such that

$$|p_n - \tan 2^{\hat{n}} \delta| < \frac{1}{2^{\hat{n}}}$$

Proofs are left as an exercise for the reader.

C. E. SHANNON

Have $[x]$ mean, as usual, the largest integer in x .
The value of λ is $\frac{1}{2}$.
Theorem 1. There exists a unique real positive number $p < \frac{1}{2}$ such that the n th prime p_n is given by
$$p_n = [2^{2^{n-1}} p] - 2^{2^{n-1}} [2^{2^{n-1}} p]$$

Note the improvement over previous results - this formula gives all the primes, not just some of them.
For anyone who finds the bracket symbol a little suspect, we have the following:
Theorem 2. There exists a real number p such that $2^{2^{n-1}} p$ is positive or negative according as n is prime or composite.

6/3/53

L84J

COVER SHEET FOR TECHNICAL MEMORANDUM

SUBJECT: The Relay Circuit Synthesizer - Case 20878

COPIES TO:

CASE FILE (HWB-WOB-JBF)(BDH)

DATE FILE

AREA CENTRAL FILES (4)

- 1 - M. L. Almquist
- 2 - H. W. Bode
- 3 - R. Bown
- 4 - E. Bruce
- 5 - A. J. Busch
- 6 - A. B. Clark
- 7 - W. H. Doherty
- 8 - E. B. Ferrell
- 9 - J. B. Fisk
- 10 - H. T. Friis
- 11 - T. C. Fry
- 12 - G. W. Gilman
- 13 - D. W. Hagelbarger
- 14 - B. D. Holbrook
- 15 - A. C. Keller
- 16 - F. A. Korn
- 17 - W. D. Lewis

MM- 53-140-52

53-180-52

DATE November 30, 1953

AUTHOR C. E. Shannon
E. F. Moore

FILING SUBJECT
(TO BE ASSIGNED BY AUTHOR)

Switching Theory

- 18 - C. A. Lovell
- 19 - M. B. McDavitt
- 20 - J. Meszar
- 21 - R. K. Potter
- 22 - F. J. Singer
- 23 - S. H. Washburn
- 24 - I. G. Wilson

THIS COPY FOR

ABSTRACT

The Relay Circuit Synthesizer is a machine to aid in switching circuit design. It is capable of designing two terminal circuits involving up to four relays in a few minutes. The solutions are usually minimal. The machine, its operation, characteristics and circuits are described.

The Relay Circuit Synthesizer - Case 20878

MM-53-140-52

MM-53-180-52

November 30, 1953

MEMORANDUM FOR FILE

Purpose and Operation

The Relay Circuit Synthesizer (Photograph 214142) is a machine to aid in the design of a certain class of relay circuits. The type of circuits it handles are two-terminal switching circuits involving up to four relays or (by simple alterations) other two-valued elements. The desired characteristics of the circuit to be designed are entered in a set of sixteen three-position switches on the front panel of the machine. After a period of computation, averaging about five minutes, the machine stops and displays a circuit satisfying the requirements. The circuit is displayed in geometric form on a card in an associated card display mechanism (Photograph 214140). The labels of the contacts on this card must, however, be interpreted in accordance with indicating lights on the front panel of the machine to obtain the proper answer to the design problem.

In about eighty per cent of the possible problems that can be set up on the machine, the solution it gives will be minimal in contacts, i.e., the number of contacts in the circuit cannot be reduced. In the remaining twenty per cent, the designs cannot be simplified by more than one contact and may, in fact, be minimal.

The sixteen input switches correspond to the sixteen possible states of the four relays in the circuit being designed. Each of these switches has three positions labeled "open," "don't care" and "closed". If, for a given state of these relays, it is desired that the circuit be open, the corresponding switch is set in the "open" position. Similarly for the "closed" position. If it does not matter whether the circuit be open or closed in this state, the switch is set at "don't care". The Synthesizer takes advantage of any switches in the "don't care" position in attempting to reduce the number of contacts used in the final circuit. It fills in these unspecified states in such a way as to minimize contact requirements. This ability to handle partially specified switching problems is one of the main features of the Synthesizer and enables it to solve problems for which analytic methods are at present ill-adapted.

In addition to the direct circuit designing procedure outlined above, the Synthesizer is equipped with controls for other modes of operation. It may be run at low speed for demonstration purposes, it may be set up to find all the circuits in its card file satisfying the requirements (not just the one with the smallest number of contacts) and it may be used to determine various mathematical properties associated with switching functions.

By changing the paper tape and the card file used (but without any internal change within the electrical part of the machine) it can be made to solve design problems involving diode circuits instead of relay contact circuits. By a still different tape and set of cards it can minimize the number of transfers in relay circuits instead of the number of contacts. With suitable tape and card file, it can solve a variety of other similar problems.

The Synthesizer represents a first step toward machine design of switching circuits. Unfortunately, although the method used in the Synthesizer may be generalized in principle to circuits involving five or more variables, the time for solution increases at an alarming rate. With five variables it would take many thousand times as long to obtain a solution. The card file and the tape would be about two thousand times their present size and would require many man years to construct. Consequently, a direct generalization of the Synthesizer is hardly indicated, even with the high speeds available in electronic computing gear.

Speed of Solution With Random Problems

An idea of the time required for the Synthesizer to solve problems may be obtained from some tests with random settings of the input switches. Using a book of random numbers, ten sets of sixteen random binary digits were obtained. These were set up as input switch settings using 0 to mean closed and 1 open, and the time required for the machine to solve each of these problems was measured. The following table gives the results of this test.

<u>Binary Digits (Switch Settings)</u>	<u>Solution Circuit No.</u>	<u>Trans- formation</u>	<u>No. of Contacts</u>	<u>Time of Solution</u>
0 0 1 1	#279	w' w	8	4min-10sec.
0 0 0 0		x' z		
0 1 1 1		y y		
1 1 0 0		z' x		
1 0 0 1	#177	w' x	6	1min-10sec.
0 0 1 0		x y		
0 0 1 0		y z		
1 1 1 1		z w		
1 0 1 0	#306	w z	10	7min-20sec.
0 0 0 1		x' y		
1 0 0 1		y' w		
0 0 0 1		z' x		
0 0 0 1	#261	w z	10	7min-7sec.
1 0 0 0		x' w		
1 0 0 1		y y		
1 1 1 0		z' x		
1 0 1 0	#212	w x	11	9min-6sec.
0 1 1 1		x' w		
1 0 0 1		y' y		
0 1 0 0		z z		

<u>Binary Digits (Switch Settings)</u>	<u>Solution Circuit No.</u>	<u>Trans- formation</u>	<u>No. of Contacts</u>	<u>Time of Solution</u>
0 1 0 1	#187	w w	9	6min-32sec.
0 0 0 0		x' x		
1 0 1 1		y y		
1 1 1 0		z z		
0 1 0 0	# 75	w x	9	6min-10sec.
1 0 1 1		x z		
1 1 1 0		y' y		
1 1 1 0		z w		
0 0 0 0	#240	w' y	5	38sec.
0 0 1 1		x' w		
0 0 1 1		y' x		
1 0 1 1		z z		
0 1 0 0	#193	w z	8	4min-30sec.
1 0 1 0		x' y		
1 0 1 0		y w		
1 1 1 0		z x		
1 0 0 0	# 84	w x	9	5min-50sec.
0 1 1 1		x' z		
1 0 1 1		y w		
1 0 1 1		z y		

The Solution Circuit Number refers to the Table in MM-52-180-45, E. F. Moore, "A Table of Four Relay Two Terminal Contact Networks". The Transformation indicates the required change of variables in interpreting the numbered circuit of this Table. The average solution time for these ten completely specified random functions was 5 min.-15 sec., and the average number of contacts in the solution was 8.5.

A second test was run with partially specified random functions. Again using the Table of Random Numbers, four switches were chosen at random for "don't care" settings; the remaining switches being given random "open" or "closed" settings. This was done four times, leading to the following results:

<u>Binary Digits (Switch Settings) D=Don't Care</u>	<u>Solution Circuit No.</u>	<u>Trans- formation</u>	<u>No. of Contacts</u>	<u>Time of Solution</u>
D 1 0 1	#334	w w	6	3min-5sec.
0 D 0 0		x x		
D 1 0 D		y y		
0 0 0 0		z z		
D 0 1 D	#189	w' w	7	6min-30sec.
D 1 0 1		x z		
0 1 0 0		y y		
D 0 1 0		z x		
0 1 0 D	#178	w y	8	7min-25sec.
0 D 1 1		x' w		
D 0 0 1		y z		
D 0 1 1		z' x		
0 0 1 D	# 58	w y	3	12sec.
0 D D 1		x w		
D 0 1 1		y' z		
1 0 1 1		z' x		

The average time of solution for these problems with four unspecified states was 4 min.-20 sec., with an average of 6 contacts.

Finally, a test was run with random problems having eight unspecified ("don't care") states. These results were as follows:

Binary Digits (Switch Settings) D=Don't Care	Solution Circuit No.	Trans- formation	No. of Contacts	Time of Solution
0 D 1 D	#204	w w	4	55sec.
D D D 0		x z		
D 0 0 1		y y		
D 0 0 D		z x		
0 D D 1	#179	w y	6	2min-55sec.
1 D 1 D		x x		
1 0 1 0		y' z		
D D D D		z' z		
0 0 D D	# 58	w y	3	40sec.
0 0 D 1		x x		
1 D D 1		y w		
0 D D D		z z		
D D D D	# 79	w' y	5	3min-15sec.
1 1 D D		x' z		
D 1 1 0		y x		
D 1 0 1		z w		

The average solution time here was 1 min.-56 sec. and the average number of contacts 4.5.

The following table summarizes these average figures:

	<u>Completely specified</u>	<u>Unspecified in 4 states</u>	<u>Unspecified in 8 states</u>
average time	5min-15sec	4min-20sec	1min-56sec
average number of contacts	8.5	6	4.5

With still more "don't care" states the solution time and average number of contacts would undoubtedly decrease still further.

General Theory of Operation

The Relay Synthesizer deals with Boolean functions of four variables. Each of the variables has two possible values, 0 to 1; in conjunction there are $2^4 = 16$ sets of values or "states" of the variables. For each of these states, a function of these variables can be either 0 to 1. Thus there are $2^{16} = 65,536$ different Boolean functions of four variables. It is known that these 65,536 functions can be subdivided into 402 classes or "types" of functions. Two functions are said to be of the same type if one may be obtained from the other by negating some of the variables or permuting some of the variables or both. Thus the function

$$w + x'(y+z)$$

is of the same type as

$$x' + z(w+y')$$

or

$$w' + y(x'+z').$$

All functions of a given type present substantially the same design problem. If a good circuit is found for one of them, it applies equally to all other functions of the same type, for it is necessary only to relabel contacts properly and it will represent these other functions.

In the memorandum referred to above, circuits are given for these 402 types of functions. At present writing, 331 of these have been proved to be minimal in contacts; the remaining 71 are known to be within one contact of being minimal. This catalog of circuits is a key part of the design procedure in the Relay Synthesizer.

The reader may wonder why the Synthesizer is necessary for designing circuits when such a catalog is available. Why not merely find the circuit corresponding to the desired function in the catalog? The answer is that it is not at all easy to find the type or class to which a given function belongs even when the function is completely specified. If the desired function is not completely specified (has one or more "don't care" states) there will in general be many types of functions consistent with the requirements, and it becomes extremely difficult to locate these in the catalog. The Synthesizer is, in fact, a machine for determining the type of a fully specified function and (in the partially specified case) the possible type having the least number of contacts in its catalog circuit.

A block diagram of the Synthesizer is shown in Figure 1, and indicates the main functional organization. The specifications of the desired circuit are set up on the input switches in the right-hand box. The catalog of the 402 types of functions appears on a paper tape in the left-hand Tape Input box. Each function occupies six lines of tape. The first four lines give the states for which the function is closed. The fifth line gives the number (in binary form) of closed states for the function, and the sixth line contains a special hole marking the end of data relating to this function, i.e., it acts as a punctuation mark separating functions on the tape.

In solving a particular problem, the tape functions are studied one by one in the machine. All permutations and negations of a particular tape function are compared with the desired specifications as set up on the input switches. When an exact match is found the machine stops, and the tape function together with the permutation being applied to it represent a solution to the problem.

In the block diagram this is carried out as follows: The tape function is stored in the memory relays. A permuting-negating network applies the equivalent of the various possible permutation and negation operations to these data. The results of each permutation-negation operation are compared with the input switches in a comparison circuit to see if a match has

occurred. If not, an error signal is fed back to the permutation sequencer, causing it to advance to the next permutation operation which is, in turn, compared, etc., until all of the 384 possible permutations and negations have been tested. Because of short-cut circuits to be described later, the machine frequently skips many of these, reducing the solution time considerably.

When the set of operations on a particular function is exhausted, the permutation sequencer sends a signal back to the tape driving circuit, and the next function is read into the memory for test. This signal also causes the card display device to drop another card from its stack. The card displayed always corresponds to the function being tested in the machine and shows the most efficient known circuit for that function.

The permutation indicator is controlled by the permutation sequencer and indicates in lights the permutation currently being tested. When the machine stops at a solution, these lights show what permutation and negation must be applied to the circuit on the card to solve the problem at hand.

In the problems involving "don't cares," the Synthesizer could be used to successively find all of the solution, but to use all this information in designing a circuit, it would be necessary to compare all the circuits obtained, and see which one is preferred. Since the grounds for preferring one circuit over another has been taken to be economy of contacts, the necessity for this comparison step has been eliminated by arranging the functions on the tape in order of increasing number of contacts, so that the first solution arrived at will automatically be the preferred one. Arranging the functions on the tape in terms of any other criterion will cause the Synthesizer to design circuits based on this criterion. If, for instance, it is desired to design relay circuits using as few springs as possible, or to design diode logic circuits using as few diodes as possible, it is only necessary to arrange the functions on the tape in order of number of springs or number of diodes, respectively.

Circuit Operation

Figure 2 is the circuit diagram of the Synthesizer. The layout of subcircuits corresponds roughly to the block diagram Figure 1. We will first describe the circuit operation in the logically simplest mode of operation -- the normal mode with all short-cut circuits eliminated. In Figure 2, then, we

assume the mode of operation switch in the "Normal" position N, the relay Q operated (eliminates permutation short cuts) and the number of state switches M are set at "Normal".

Since the Synthesizer is essentially a closed loop system, it is difficult to find a point at which to start a description of its operation. It is perhaps simplest to assume that the machine has just finished testing one function on the tape. The relay H may then be assumed to have just operated locking in to the make on R_s , since the tape reader will be at the division line between functions and consequently R_s operated. Operation of H releases the hold on the memory relays (M_0, M_1, \dots, M_{15}) and also the hold on the steering counter relays ($W_{M1}, Z_{M1}, W_{M2}, Z_{M2}, W_{M3}, Z_{M3}$), thus resetting this counter to zero. It also applies voltage to the teletype magnet which, a moment later, will pull free of the tape and hence release R_s . This releases H and reconnects the holds of the steering counter and the memory relays. It also establishes a path to the slow relay SO through its own back contact SO'. SO now acts like a slow buzzer, producing pulses at a rate of about six per second and relay U follows these pulses through the SO make contact.

The pulses produced by U operate the teletype magnet, advancing it line by line until it reaches the line with an R_s hole, at which point the back contacts on R_s open both the buzzer circuit to SO and the teletype magnet circuit through U. The pulses produced by U are also fed into the three-stage binary counter consisting of three WZ pulse dividers $W_{M1}Z_{M1}$, $W_{M2}Z_{M2}$, $W_{M3}Z_{M3}$. This counter, therefore, keeps track of the line of tape, counting from the last division between two tape functions (R_s hole). This counter controls the steering trees leading into the memory relays M_0, M_1, \dots, M_{15} and the number of state relays V_1, V_2, V_4, V_8 . The first line of tape after the R_s line is fed into M_0, M_1, M_2, M_3 , the second line into M_4, M_5, M_6, M_7 , the third into M_8, M_9, M_{10}, M_{11} , the fourth into $M_{12}, M_{13}, M_{14}, M_{15}$, and the fifth into V_1, V_2, V_4, V_8 . A section of the tape is shown in Figure 3.

The completion of this tape reading operation, indicated by closure of R_s , puts ground on lead 106 leading into the permutation-negation network.

Permuting and Negating Circuits

These circuits enable the machine to apply the 384 negation and permutation operations to the tape function stored in the memory to compare it with the desired function set on the input switches.

The negation-permutation sequencer consists of nine WZ pairs connected in a form of counting circuit which can go through 384 different states. Starting from the high speed (pulsed) end of this circuit, the first six WZ pairs, E, D, B, C and A, relate to permutations and can go through twenty-four states corresponding to the $4! = 24$ permutations of the four variables. The other four stages w, x, y, z relate to negating the variables and can go through sixteen states corresponding to the sixteen ways of negating four variables. In combination this gives 384 states. E.V.

In the circuit, imagine Q operated, F_0 and F_{16} released and that F_3 is pulsed, so that a series of pulses is applied to line 109. The negation-permutation sequencer will then proceed through the 384 negation-permutation operations. This sequence is shown in the accompanying Table I for the first twenty-four of these, i.e., a full set of permutations. At the twenty-fourth step this sequence repeats for the permutation relays but a pulse is applied at lead 250, advancing the negating relays one step. The negating relays go through the sequence shown in Table II, advancing one step after the permuting relays have gone through a full set of permutations. In this manner the full set of 16×24 combinations is exhausted.

Y W X Z	0 1 1 0 1	1 0 0 1 0	11
W X Y Z	1 1 1 0 0	0 0 0 1 1	12
W X Z Y	0 0 1 0 0	1 1 0 1 1	13
W X Y Z	1 0 1 0 0	0 1 0 1 1	14
W Z X Y	0 0 0 0 0	1 1 1 1 1	15
W Y X Z	1 0 0 0 0	0 1 1 1 1	16
W Y Z X	0 1 1 0 0	1 0 0 1 1	17
Y Z W X	1 1 1 1 0	0 0 0 0 1	18
Z X W Y	0 0 1 1 0	1 1 0 0 1	19
Y X W Z	1 0 1 1 0	0 1 0 0 1	20
X Z W Y	0 0 0 1 0	1 1 1 0 1	21
X Y W Z	1 0 0 1 0	0 1 1 0 1	22
Z Y W X	0 1 1 1 0	1 0 0 0 1	23

Table I

Sequence of Permutations

	Relays					Relays					Permutation			
	W _A	W _B	W _C	W _D	W _E	A	B	C	D	E	W	X	Y	Z
	(1 means operated)										Becomes			
0	0	0	0	0	0	1	1	1	1	1	W	X	Y	Z
1	0	0	0	1	1	1	1	1	0	0	W	Y	Z	X
2	0	0	0	1	0	1	1	1	0	1	W	Z	Y	X
3	0	0	1	1	1	1	1	0	0	0	W	Y	X	Z
4	0	0	1	1	0	1	1	0	0	1	W	Z	X	Y
5	0	0	0	0	1	1	1	1	1	0	W	X	Z	Y
6	0	1	0	0	0	1	0	1	1	1	Y	X	W	Z
7	0	1	0	1	1	1	0	1	0	0	Z	Y	W	X
8	0	1	0	1	0	1	0	1	0	1	Y	Z	W	X
9	0	1	1	1	1	1	0	0	0	0	X	Y	W	Z
10	0	1	1	1	0	1	0	0	0	1	X	Z	W	Y
11	0	1	0	0	1	1	0	1	1	0	Z	X	W	Y
12	1	1	0	0	0	0	0	1	1	1	X	Y	Z	W
13	1	1	0	1	1	0	0	1	0	0	Y	Z	X	W
14	1	1	0	1	0	0	0	1	0	1	Z	Y	X	W
15	1	1	1	1	1	0	0	0	0	0	Y	X	Z	W
16	1	1	1	1	0	0	0	0	0	1	Z	X	Y	W
17	1	1	0	0	1	0	0	1	1	0	X	Z	Y	W
18	1	0	0	0	0	0	1	1	1	1	X	W	Z	Y
19	1	0	0	1	1	0	1	1	0	0	Y	W	X	Z
20	1	0	0	1	0	0	1	1	0	1	Z	W	X	Y
21	1	0	1	1	1	0	1	0	0	0	Y	W	Z	X
22	1	0	1	1	0	0	1	0	0	1	Z	W	Y	X
23	1	0	0	0	1	0	1	1	1	0	X	W	Y	Z

Table II
Sequence of Negations

Relays				Relays				Variables			
W _w	W _x	W _y	W _z	W	X	Y	Z	W	X	Y	Z
								Become			
0	0	0	0	1	1	1	1	W	X	Y	Z
0	0	0	1	1	1	1	0	W	X	Y	Z'
0	0	1	1	1	1	0	0	W	X	Y'	Z'
0	0	1	0	1	1	0	1	W	X	Y'	Z
0	1	0	0	1	0	1	1	W	X'	Y	Z
0	1	0	1	1	0	1	0	W	X'	Y	Z'
0	1	1	1	1	0	0	0	W	X'	Y'	Z'
0	1	1	0	1	0	0	1	W	X'	Y'	Z
1	1	0	0	0	0	1	1	W'	X'	Y	Z
1	1	0	1	0	0	1	0	W'	X'	Y	Z'
1	1	1	1	0	0	0	0	W'	X'	Y'	Z'
1	1	1	0	0	0	0	1	W'	X'	Y'	Z
1	0	0	0	0	1	1	1	W'	X	Y	Z
1	0	0	1	0	1	1	0	W'	X	Y	Z'
1	0	1	1	0	1	0	0	W'	X	Y'	Z'
1	0	1	0	0	1	0	1	W'	X	Y'	Z

At the end of this sequence, a ground is applied to line 135 which initiates reading in a new function.

It may also be noted that if relay Q is released and Fl6 is operated a ground is applied directly to line 250, the input to the negating part of the counter. This will

cause the counter to skip a set of permutations and advance directly in the negating sequence by one step. Operation of F_{16} also releases the plus side of the permutation relays in the sequencer, resetting them to zero. The function of F_{16} is to shortcut some of the calculation in certain cases as will be described later.

In a similar way, operation of F_9 with Q released advances the W_A and W_B parts of the permutation sequence by one step, skipping a subset of six permutations in which W_C , W_D and W_E take part. F_9 releases the plus to these three WZ pairs, resetting them to zero. This also is used for shortcut purposes.

The permuting and negating relays A, B, C, D, E and W, X, Y, Z are operated from back contacts of the corresponding W relays in the WZ pairs of the sequencer. Thus they assume the complementary states as shown in Tables I and II. The function of these nine sets of relays is to interchange sixteen leads representing the function in the memory relays in accordance with the permutation and negation in the sequencer.

The logical organization of this circuit can be represented in a symbolic form by Figure 4, which indicates the effect of the negating and permuting relays on the variables of the tape function, (not the effect on the sixteen leads). Thus, the W relay negates the variable W when released, the X relay negates X, etc. The A relay interchanges W and X and also Y and Z when released, the B relay interchanges the variables now appearing (after the possible A interchange) on the first and third lines, etc. It will be found that the twenty-four combinations of A, B, C, D, and E produced by the sequencer (Table I) lead to the twenty-four permutations of the four variables as shown in Table I.

Now the circuit does not work with the four Boolean variables but with sixteen lines representing the sixteen states of the four variables. Negating a variable, say W, corresponds to interchanging the eight lines (or states) for which W is 1 with the corresponding eight lines for which W is zero. Thus in the premuting circuit, the W negation box of Figure 4 becomes eight reversing or interchanging circuits operated by the relays W_1 , W_2 , W_3 , W_4 . A similar statement applies to the negation of the other variables and the permuting of the variables by the A, B, C, D and E relays.

To summarize, the sequencer can go through 384 states representing the 384 permutations and negations. The negating-permuting network sets up the corresponding interchanges of the sixteen lines from the memory to the input switches. At the memory end, these lines are given plus or minus voltage according as the memory function is open or closed. At the input switch end, after the permutation and negation, these voltages are compared with the settings of the input switches.

There are two types of comparison circuits. The first type, Figure 5, applies to switches 0, 7, 8 and 15. It will be seen that F_0 will operate if the lead from the permuting network is positive and the switch is set at "closed," or if the lead is negative and the switch is set at "open," i.e., if there is a disagreement between the switch setting and the value coming in from the permuting network. If the switch is set at "don't care," F_0 will not operate. It will also be seen that the red and green lights will indicate "closed" and "open" settings of the switch respectively, while if set at "don't care" the red or green light will indicate minus or plus coming in from the permuting network.

The comparison circuit for the other switches is somewhat different. There are two relays F_1 and F_2 common to all the other switches. If a particular switch is set at "closed," the line from the permuter goes through a diode to F_1 , the other side of F_1 being minus (when the test is made). Thus F_1 will operate if a plus appears on the line from the permuter (disagreeing with the "closed" position of the switch). If the switch is set at "open," the path from the permuter goes through the same diode but in the opposite direction to F_2 , whose other side is connected to plus. Hence F_2 will operate if a minus comes in from the permuter. The red and green lamps are connected substantially as before.

Returning now to the description of the operating sequences in the machine, we recall that the completion of tape reading of a function into the memory was signified by closure of R_0 . This applies ground at lead 106 into a long "equality chain" of contacts. This chain is closed only if all of the W relays in the WZ pairs of the sequencer agree in position with their corresponding Z relays. This being true, ground is applied to the permuting and negating network, and, as already described, one or more of the F relays ($F_0, F_7, F_8, F_{15}, F_1, F_2$) will operate unless the tape function as permuted through the network agrees with the input

function. Assuming there is a disagreement, one at least of F_3 , F_9 , F_{16} will operate, grounding the input to the negation-permutation sequencer. This advances the W relays of the sequencer one step in the sequence, and causes a disagreement between at least one of the W relays and its corresponding Z relay in the WZ pairs. This disagreement, in turn, opens the "equality chain," releasing the F relays which, in turn, removes the ground from the sequencer and allows its Z relays to follow their corresponding W relays. When equality has again been established, ground is again applied through the "equality chain" to the permuting network and the next permutation of the sequence (now set up on the permuting network) is tested in the same way. This cycle of operations continues until the full set of permutations and negations has been tested. After the last permutation, the next ground goes through a Z_w contact and the mode of operation switch to operate H, signifying the completion of tests on the current function and initiating reading the tape for the next function as previously described.

If, at some point, the permuted tape function matches the input function, no F relay will operate and the cycle is stopped. Relay J will operate and, in turn, L through the chain of back contacts on the F relays. The operation of L rings the gong indicating a solution, and pulses the message register for counting purposes.

Short-Cut Operation

We now describe the short-cut provisions. If the short-cut eliminator is "off," relay Q will release, rearranging the inputs to the sequencer. In the permuting network it will be seen that the lines on the zero level and on the 15 level are not switched after the vertical column of Z contacts, i.e., after emerging from the negating part of this circuit. This means that if a disagreement occurs on either of these lines, it will persist throughout all the permutations, which only change the switches A, B, C, D and E in this network. Hence, in case of such a disagreement it is not necessary to test all of these permutations but the machine can proceed immediately to the next negation saving a great deal of time.

In the circuit, when Q is released, operation of F_0 or F_{15} pulses directly into the negating part of the sequencer and resets the permuting part to zero.

In a similar manner, it will be seen that the lines at the 7 and 8 level in the permuter are not switched after the B contacts. This means that a disagreement on either of these lines, indicated by operation of F_7 or F_8 , will persist over the subset of six permutations in which C, D and E change. Hence it is unnecessary in such a case to test each of these individually and the machine advances to the next permutation involving a change of A or B. In the sequencer, a ground is applied at the input to the A, B stages and C, D, E stages are reset to zero. This is done by relay F_9 which will operate if either F_7 or F_8 indicates disagreement.

One further short-cutting device has been incorporated in the machine. With each tape function is included, in binary form, the number of states for which that function is closed. As previously described, this number is stored in the relays $V_1, V_2, V_4, V_8, V_{16}$ when the function is read off the tape. On the front panel of the machine are two seventeen-point switches labeled Max and Min. The Min switch should be set at a number equal to the number of input switches in the "closed" position. The Max switch should be set at this number plus the number of "don't cares". Now, regardless of how the "don't cares" may be filled in, the number of closed states will be within this range (including the end points). A function from the tape could not possibly be satisfactory unless its number of states lies within this range. The machine is arranged to compare these numbers and, if this condition is not satisfied, to skip the function completely and go immediately to the next function on the tape.

The comparison is carried out in the "number of states comparison circuit". The contacts on the V relays are arranged in the topological dual of an ordinary tree. This implies that if the number n is registered (in binary form) in the V relays, then all of the vertical leads labeled zero to n at the Min switch will be connected together, but the two groups are not connected. It will be seen, therefore, that if the number on the V switches lies in the range covered by the Max and Min settings, then the Max and Min swingers will not be connected. If the V number is outside this range then the Max and Min swingers will be connected. If the Max and Min swingers are connected, the operation of R_5 closes a path to operate H and start reading in a new function immediately.

It is necessary to use five relays - V_1, V_2, V_4, V_8 , and V_{16} - to represent all of the numbers from 0 to 16 inclusive, but there were only four holes readily available on

the tape for reading into these relays. Consequently four of the relays are read into directly through the steering relays, and a special artifice is used to get the fifth digit stored in V_{16} .

Since the only case in which this digit equals 1 is when the number of states is 16, and all the other four relays are released, this relay is operated through the back contacts of V_1 , V_2 , V_4 , and V_8 in series. But since V_1 , V_2 , V_4 , and V_8 are also all released when the number of states is 0, a contact of M_0 is also included in the operate path, to distinguish between these two cases.

Without the short-cutting features the average time of solution for a completely specified function would be over an hour; with short cuts it is about five minutes.

Indicating Circuits

A set of indicating lights is provided which shows the permutation and negation that must be applied to the tape function (when a solution has been found) to transform it into the function on the input switches. The eight negating lights are connected in a simple fashion to the W, X, Y and Z coils. If the W relay is out, for example, the W' lamp lights up by a current through the W coil (not sufficient to operate the W relay). If the W relay is operated, the W lamp lights up by current through the W_w contact.

The circuit for the permuting part is more complex. However, on tracing through the circuits it will be found that the lights always receive proper voltages to indicate the permutation set up on the A, B, C, D, E relays. For example, in the first (identity) permutation A, B, C, D and E are all operated. It will be seen that the eight center points between pairs of lamps receive the following voltages: (0 indicates floating)

+ - 0 0
0 0 + - .

Hence the diagonal series of lamps

W - - - -
 - X - - - -
 - - Y -
 - - - Z

will be lighted. Note that the lamps connected to floating points receive half voltage by a sneak path through the two lamps in series. This is not sufficient to illuminate them perceptibly.

Another permutation indicating light circuit has been provided for trouble shooting and for better observation of the machine while in action. This consists of twenty-five small neon lamps. Twenty-four of these correspond to the twenty-four permutations of the variables. These are arranged in a rectangle six wide and four high. In operation without short cuts, these lamps light sequentially from left to right across the first row, then across the second, etc. In short cuts due to the F_0 and F_{15} relays the whole pattern of twenty-four permutations is skipped. In short cuts due to F_7 and F_8 a horizontal row in this display is skipped (only the first lamp of the row going on).

The circuit controlling these lights consists of a tree on relays A and B which selects the row and a second tree on C, D and E which selects the column. Only the lamp at the intersection point will go on. Sneak paths through other lamps all involve at least three lamps in series and the voltage is not sufficient for breakdown of such a series combination.

The twenty-fifth lamp is connected to light up if the C, D and E relays get into either of the two other possible states which do not correspond to permutations in the regular sequence of operations. It can thus indicate certain trouble conditions.

Other Modes of Operation

With the mode of operation switch set in the P position (periodic), the machine does not advance the tape after the sequence of permutations and negations but periodically goes through the tests on the function in the memory. In this switch position the path to the H relay, which ordinarily initiates the tape reading process, is open. This mode is sometimes useful for trouble shooting.

In the S position (step-by-step), the machine tests a permutation and then stops until the Run switch is operated and released. The path which normally puts ground on the relays F_3 , F_9 , F_{16} is opened and replaced by a contact on the Run switch connected to a condenser. When the Run switch is off, this condenser charges, and when pressed for a step in the operation it discharges through F_3 , F_9 or F_{16} . Only enough charge is stored to operate these relays once. For the next step the Run switch must be released and pressed again.

In the L mode (low-speed), the machine operates as in the normal mode except at a much lower speed. This is achieved in a fashion similar to the step-by-step operation but with the function of the Run switch replaced by relay N. The N relay is operated by the G relay which is connected in a relaxation oscillator circuit using a gas tube. The condensers charge up sufficiently to break down the gas tube which operates G, closing its make contact and discharging the condenser which then starts recharging. This slow oscillation of G causes N to oscillate slowly which, in turn, allows the solution to proceed at a slow rate.

In Mode Q (self-restarting), the machine does not stop at a solution but rings the gong, pulses the message register, and then proceeds to the next permutation or negation in the sequence. When a solution is reached in this mode, the operation of relay L causes the message register to operate. This releases relay K which releases the message register and also applies voltage to slow-operate relay G. Operation of G energizes N, which in turn advances the permutation sequencer one step and also energizes K. K locks in releasing G and in turn, N, and the solution proceeds. This mode of operation can be used to find all of the solutions to the given problem, rather than just the first one.

C. E. SHANNON

E. F. MOORE

Att:
Appendices A and B
Photographs 214140 through 214143
Figures 1 through 5

Appendix A

Main Components and Their Functions

Relays and Other Electromagnetic Components

M_0, M_1, \dots, M_{15}

Memory relays. These register the values of the function read off the tape for its sixteen possible states. If M_i is operated, the function is closed in state i .

W_1, W_2, W_3, W_4

Four parallel relays (to give sufficient contacts). These relays negate the variable W of the tape function. This is done in the negating and permuting network by interchanging the eight leads corresponding to the variable $W=1$ with the corresponding eight leads for which the variable W is zero.

X_1, X_2, X_3, X_4

Similar negating relays for the variable X .

Y_1, Y_2, Y_3, Y_4

Similar negating relays for the variable Y .

Z_1, Z_2, Z_3, Z_4

Similar negating relays for the variable Z .

A_1, A_2, A_3, A_4

B_1, B_2, B_3, B_4

C_1, C_2, C_3, C_4

D_1, D_2, D_3, D_4

E_1, E_2, E_3, E_4

Permuting relays. The function of these relays is to permute the sixteen lines from the memory relays according to the various permutation of the variables W, X, Y and Z in the tape function. By suitable combinations of operation and release of these five sets of relays, the interchanges corresponding to any of the twenty-four permutations are possible.

$W_z Z_w, W_x Z_x, W_y Z_y$

$W_z Z_z, W_a Z_a, W_b Z_b$

$W_c Z_c, W_d Z_d, W_e Z_e$

WZ relays arranged in a counting circuit to go through the 384 permutations and negations applied to the sixteen leads in the permuter. These WZ pairs control the preceding W, X, \dots, E relays, thus W_1, W_2, W_3, W_4 are controlled by the W_w relay of the $W_w Z_w$ pair.

Appendix A (Continued)

F_0, F_7, F_8, F_{15}

Failure relays. Operation of F_0 , for example, corresponds to failure of the permuted line coming into switch 0 to match the value on input switch I_0 . Operation of a failure relay causes the machine to proceed to try another permutation or tape function.

F_1, F_2

These are failure relays which are operated by a failure to match on any of the other switches not taken care of specifically by F_0, F_7, F_8 or F_{15} .

F_3, F_9, F_{16}

Secondary failure relays. These are operated by the preceding failure relays and sort out the type of short cut (if any) available. F_{16} causes the permuter to advance to the next negation (skipping all permutations of the current negation). F_9 causes the permuter to skip the current subset of six permutations out of the twenty-four, advancing the AB part of the permutation one unit. F_3 causes an advance of only one in the permutation.

R_0, R_1, R_2, R_3, R_s

These relays are controlled by the five fingers of the tape reading mechanism. For example, a hole in the 2 row of the tape operates R_2 . R_0, R_1, R_2, R_3 carry information to the memory relays M_0, \dots, M_{15} and also to the number of state relays V_1, V_2, V_4, V_8 . R_s marks the end of data relating to one function on the tape.

S_1, S_2, S_3, S_4

Steering relays. These relays steer, by means of four trees, the tape readings on R_0, R_1, R_2, R_3 into the memory relays and the number of state relays V_1, V_2, V_4, V_8 .

Appendix A (Continued)

$W_{m1}Z_{m1}, W_{m2}Z_{m2}, W_{m3}Z_{m3}$

WZ pairs controlling the steering relays S_1, S_2, S_3, S_4 . These pairs sequence the steering for successive lines of tape into the appropriate memory and number of state relays.

$V_1, V_2, V_4, V_8, V_{16}$

Number of state relays. These relays register in binary form the number of states for which the function currently in the memory relays is closed.

W_sZ_s

A WZ pair for operating the card display unit. It causes successive functions on the tape to operate alternately the right and left solenoids S_r and S_l of the display unit.

S_r, S_l

Right and left solenoids of the display unit for releasing cards one by one from the stack.

H

End-of-permutations relay. This operates when the machine has tested all permutations of the current tape function, and initiates analysis of the next function on the tape.

L

Success relay. This operates when the machine finds a solution to the problem.

Q

Short cut eliminator. When operated, this relay eliminates short cuts in the permutation sequence.

J

A delaying and checking relay in the basic closed loop of the system. J operates when all of the WZ pairs in the permutation counter are in agreement.

SO

Slow-operate relay in a buzzer circuit for producing pulses to step the tape via relay U.

U

Secondary relay operated by SO.

Appendix A (Continued)

G	Reed relay in a slow relaxation oscillator circuit for controlling low-speed operation via secondary relay N.
N	Secondary relay controlled by G.
K	Control relay relating to low-speed and self-restarting modes of operation.
MR	Message register for counting solutions to a problem.
R	A relay for connecting the 110 volt supply only when the 24 volt supply is on.
Gong	A bell operated by L which sounds when a solution is found.
Tape Reader	A five-hole teletype tape transmitter. The standard functions are arranged on tape in order of increasing numbers of contacts.

Appendix B

Manually Operated Switches

I_0, I_1, \dots, I_{15}

Problem input switches. These switches have three positions, "open," "don't care," and "closed," and are set to correspond to the desired characteristics of the circuit to be designed in its sixteen states.

MOS

Mode of operation switch. This is a five-position switch which determines the mode of operation of the machine. In clockwise order these modes are:

P = Periodic. It continues cycling through the same permutations without advancing to the next function.

Q = Step-by-step. In this mode the machine tests the permutations one at a time under control of the key switch. This switch must be pressed once for each permutation.

N = Normal operation. Runs at regular speed to the first solution and then stops.

S = Self-restarting. At each solution, it rings the gong and adds a count to the message register, and then advances to the next solution.

L = Low-speed. Similar to normal, but at low-speed for demonstration and test purposes.

SCE

Short cut eliminator. In the "On" position this switch operates relay Q and eliminates short cuts in the permuting sequence.

NF

Next function button. Pressing this pushbutton operates relay H, causing the machine to advance to the next function on the tape, omitting any remaining permutations of the current function.

Appendix B. (Continued)

Manually Operated Switches

Run

Starts the machine operating by closing its fundamental operating feedback loop.

On-Off

Turns power on for the machine.

Max, Min

Both of these switches have seventeen points labeled, 0, 1, 2, ..., 16; the Min switch has an additional point labeled "Normal". In use, the Min switch is set at the number of states for which the function to be designed is closed. The Max switch is set at this number plus the number of "don't care" states. The machine then skips functions from the tape whose number of closed states do not lie in this range, thus shortening the solution time. If the Min switch is set at "Normal" this shortening feature is eliminated.

BOUNDS ON THE DERIVATIVES AND RISE TIME
OF A BAND AND AMPLITUDE LIMITED SIGNAL

April 8, 1954

Both experience and intuition suggest that a function of time $f(t)$ which is bounded in amplitude range ($|f(t)| \leq A$) and in bandwidth (the spectrum vanishes for angular frequencies greater than ω_0) has bounded slope, a bounded second derivative, etc., and that there is a certain minimum time required to go from a maximum negative to a maximum positive amplitude. Indeed, one feels that the maximum slopes, and higher derivatives, and the fastest rise times will occur with a sine wave having the highest allowed amplitude and the highest allowed frequency.

This note establishes some theorems of this general sort.

Theorem I: Let the function $f(t)$, of integrable square, be both amplitude limited and band limited:

$$\begin{aligned} |f(t)| &\leq A & \text{all } t \\ F(\omega) &= 0 & |\omega| > \omega_0 \end{aligned}$$

where $F(\omega)$ is the Fourier transform of $f(t)$. Then

$$\begin{aligned} f'(t) &\leq A\omega_0 & \text{all } t \\ f''(t) &\leq A\omega_0^2 & " \\ \dots &\dots & " \\ f^n(t) &\leq A\omega_0^n & " \\ \dots &\dots & " \end{aligned}$$

Proof: If we can prove the theorem for a particular t , it will follow for all t , since we can shift $f(t)$ along the time axis without affecting the assumptions of the theorem or its conclusions. We will prove the theorem for the particular time $t_1 = \frac{\pi}{2\omega_0}$. Now apply the sampling theorem of $f(t)$, expanding it in terms of its samples:

$$f(t) = \sum_{-\infty}^{\infty} a_n \frac{\sin \omega_0 t}{\omega_0 t - n\pi}$$

$$f'(t) = \sum_{-\infty}^{\infty} a_n \frac{[\omega_0(\omega_0 t - n\pi) \cos \omega_0 t - \omega_0 \sin \omega_0 t]}{(\omega_0 t - n\pi)^2}$$

$$f'(\frac{\pi}{2\omega_0}) = -\sum a_n \frac{\omega_0}{\pi^2(n-1/2)^2} \leq \sum |a_n| \frac{\omega_0}{\pi^2(n-1/2)^2}$$

since the absolute value on a_n makes all terms positive, Now a_n is the value of $f(t)$ at $t = \frac{n\pi}{\omega_0}$ and consequently $|a_n| \leq A$. Hence

$$f'(\frac{\pi}{2\omega_0}) \leq \sum A \frac{\omega_0}{\pi^2(n-1/2)^2}$$

$$= \frac{A\omega_0}{\pi^2} \sum_{-\infty}^{\infty} \frac{1}{(n-1/2)^2}$$

$$= \frac{8A\omega_0}{\pi^2} (\frac{1}{1} + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \dots)$$

$$= A\omega_0.$$

This proves the desired result for the first derivative.

The results for higher derivatives can be obtained inductively. $f'(t)$ is band-limited, of integrable square, and, as we have just shown, amplitude limited to $A\omega_0$. Hence, f'' will be amplitude limited by:

$$f''(t) \leq (A\omega_0)\omega_0 = A\omega_0^2$$

and by obvious induction

$$f^{(n)}(t) \leq A\omega_0^n$$

It will be noted that these bounds are the maximum derivatives that would be obtained for a sine wave of the highest allowed amplitude and frequency, $f(t) = A \sin \omega_0 t$. While such a wave does not satisfy our integrable square assumption, it is possible to approximate the bounds given as closely as desired by taking a sine wave of nearly top frequency and nearly top amplitude and multiplying it by a very slowly decaying function of the type $\frac{\sin kt}{kt}$ (k very small). This produces a function satisfying all the conditions with maximum derivatives approximating to the upper bounds given. Consequently these bounds are the best possible.

We now consider the problem of total rise of a function over an interval. Again we would conjecture that the shortest time for a rise from negative peak to positive peak amplitude

would be obtained by use of a sine wave of the greatest allowed frequency and amplitude and hence would be $\pi\omega_0$ seconds. We have not been able to prove a result quite this good but will show the following:

Theorem II: Under the same conditions on $f(t)$ as in Theorem I, it takes at least $3 \frac{1}{12} \omega_0$ seconds for $f(t)$ to change from $-A$ to $+A$.

Proof: We will show that if $f(0) = -A$, and $f(t_3) = +A$, then $f'(t)$ for $0 \leq t \leq t_3$ lies always under or on the curve $g(t)$ shown in Figure 1. This curve consists of five sections, a straight line segment of slope $A\omega_0^2$, a parabolic segment whose second derivative is $-A\omega_0^3$ and which is tangent to the first segment and to the third segment, a horizontal straight line at height $A\omega_0$. The last two segments are reflections of the first two. In the first place, if $f(0) = -A$, then $f'(0) = 0$, for $f(t)$ is an entire function because of the band limitations, and if $f'(0)$ were not equal to zero, $f(t)$ would run outside its amplitude limit A in the neighborhood of zero.

Now

$$\begin{aligned} f'(t) &= f'(0) + \int_0^t f''(t) dt \\ &\leq 0 + \int_0^t |f''(t)| dt \\ &\leq \int_0^t A\omega_0^2 dt = A\omega_0^2 t. \end{aligned}$$

Hence $f'(t)$ lies under or on the sloping straight line section. Also $f'(t) < A\omega_0$ so it lies under the horizontal segment. Next we show that it cannot lie in the small triangular shaped region T. Suppose in contradiction that $f'(t)$ did lie in this region, passing through a point p at $t = t_0$ as shown. At t_0 we have either (A) $f''(t_0) \geq g'(t_0)$ or (B) $f''(t_0) < g'(t_0)$.

Assume first case (A). We may write

$$f'(t_2) = f'(t_0) + (t_2 - t_0) f''(t_0) + \int_{t_0}^{t_2} \int_{t_0}^{t_2} f'''(t) dt dt. \quad (1)$$

We also have

$$g(t_2) = g(t_0) + (t_2 - t_0) g'(t_0) + \int_{t_0}^{t_2} \int_{t_2}^{t_2} g''(t) dt dt. \quad (2)$$

The three right-hand members of (1) dominate the corresponding members of (2). $f'(t_0) > g(t_0)$ since we assumed $f'(t_0)$ in the triangular region. $f''(t_0) \geq g'(t_0)$ since we are assuming case (A). $f'''(t) \geq g''(t)$ since the g curve has the greatest negative second derivative allowed by Theorem I. We conclude that $f'(t_2) > g(t_2)$, and the f' curve is over the horizontal line at t_2 , a contradiction which excludes case (A).

A similar argument applies to case (B) working backward to the point t_1 . In equations (1) and (2), read t_1

for t_2 and notice that the coefficient $(t_1 - t_0)$ now becomes negative. This allows the same argument to go through with the condition reversed on the relation of $f''(t_0)$ and $g'(t_0)$, and the resulting contradiction excludes case (B), which shows the impossibility of a curve in the triangular region.

An exactly similar argument working backward from t_3 shows that $f'(t)$ must lie under or on the right-hand sloping line and curved segment. Now if $f'(t)$ is always under $g(t)$ the area under $f'(t)$ from 0 to t_3 is less than or equal to that under $g(t)$. In order that $f(t)$ run from $-A$ to 0 to $+A$ at t_3 the area under $f'(t)$ must be at least $2A$ and hence so must that under $g(t)$. A simple integration of the $g(t)$ curve shows that this requires $t_3 \geq 3 \frac{1}{12}$. This proves the desired result.

It would no doubt be possible to improve the value $3 \frac{1}{12}$ by more elaborate arguments of the same general type, finding better $g(t)$ functions with properly banded values of $g''(t)$, $g^{iv}(t)$, etc. It seems difficult however to obtain the conjectured value by this method,

C. E. SHANNON

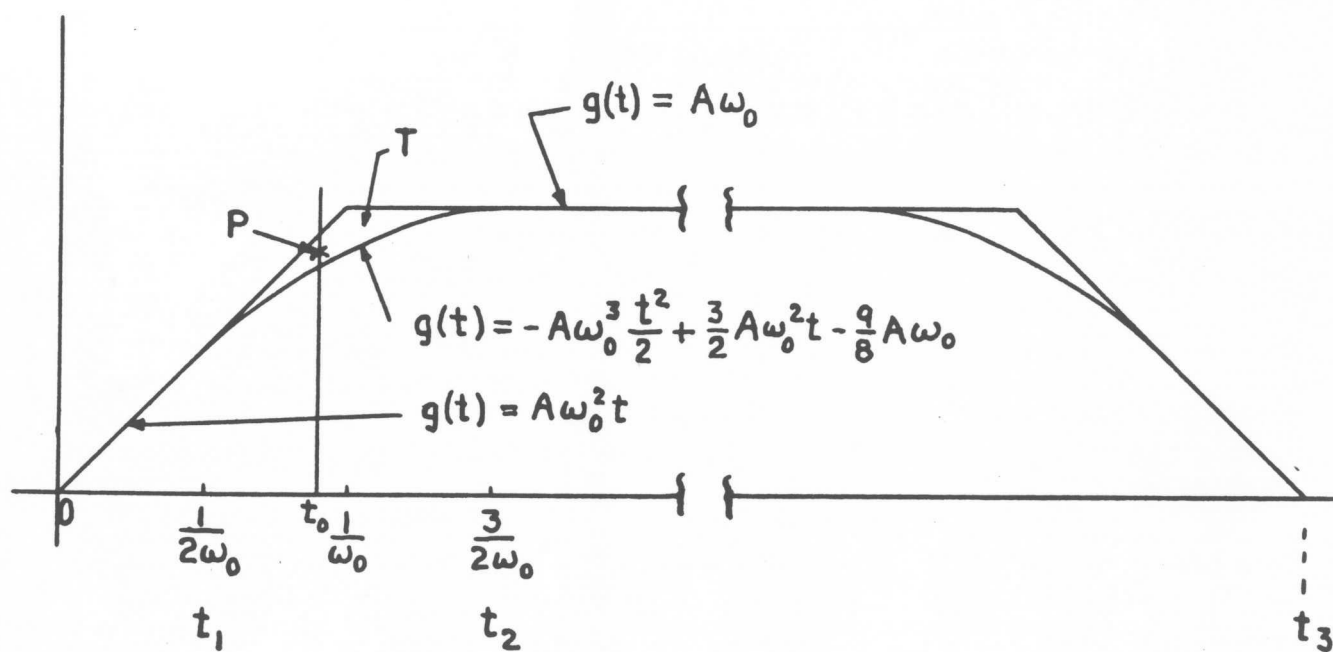


Fig. 1

COVER SHEET FOR TECHNICAL MEMORANDUM

SUBJECT: Concavity of Transmission Rate as a Function of Input Probabilities - Case 20878

COPIES TO:

CASE FILE

DATE FILE

AREA CENTRAL FILES (4)

- 1 - HWB-WOB-JBF
- 2 - H. W. Bode
- 3 - W. R. Bennett
- 4 - H. S. Black
- 5 - C. A. DeSoer
- 6 - E. N. Gilbert
- 7 - R. E. Graham
- 8 - D. W. Hagelbarger
- 9 - J. L. Kelly
- 10 - S. P. Lloyd
- 11 - L. A. MacColl
- 12 - B. McMillan
- 13 - E. F. Moore
- 14 - J. R. Pierce
- 15 - S. O. Rice
- 16 - D. Slepian

MM- 55-114-28

DATE June 8, 1955

AUTHOR C. E. Shannon

FILING SUBJECT

(TO BE ASSIGNED BY AUTHOR)

THIS COPY FOR

Information Theory

ABSTRACT

The following theorem is proved: In a discrete noisy channel without memory the rate of transmission R is a concave downward function of the probabilities P_i of the input symbols. Hence any local maximum of R will be the absolute maximum or channel capacity C .

Concavity of Transmission Rate as a Function of Input Probabilities
- Case 20878

MM-55-114-28

June 8, 1955

MEMORANDUM FOR FILE

Theorem:

In a discrete noisy channel without memory, the rate of transmission R is a concave downward function of the probabilities P_i of the input symbols. Hence, any local maximum of R will be the absolute maximum or channel capacity C .

Proof: We have

$$\begin{aligned} R &= H(y) - H_x(y) \\ &= -\sum_i Q_i \log Q_i + \sum_i P_i \alpha_i \end{aligned}$$

where the Q_i are the probabilities of the various received symbols and α_i is the conditional entropy of the received symbol when the transmitted symbol is the i -th one.

A condition for concavity of R is that $\frac{\partial^2 R}{\partial P_j \partial P_k} = R_{jk}$ be a negative semi-definite form.* We have

$$\frac{\partial R}{\partial P_j} = -\sum_i (1 + \log Q_i) p_j(i) + \alpha_j$$

using the fact that $Q_i = \sum_j P_j p_j(i)$.

$$R_{jk} = \frac{\partial^2 R}{\partial P_j \partial P_k} = -\sum_i \frac{1}{Q_i} p_j(i) p_k(i)$$

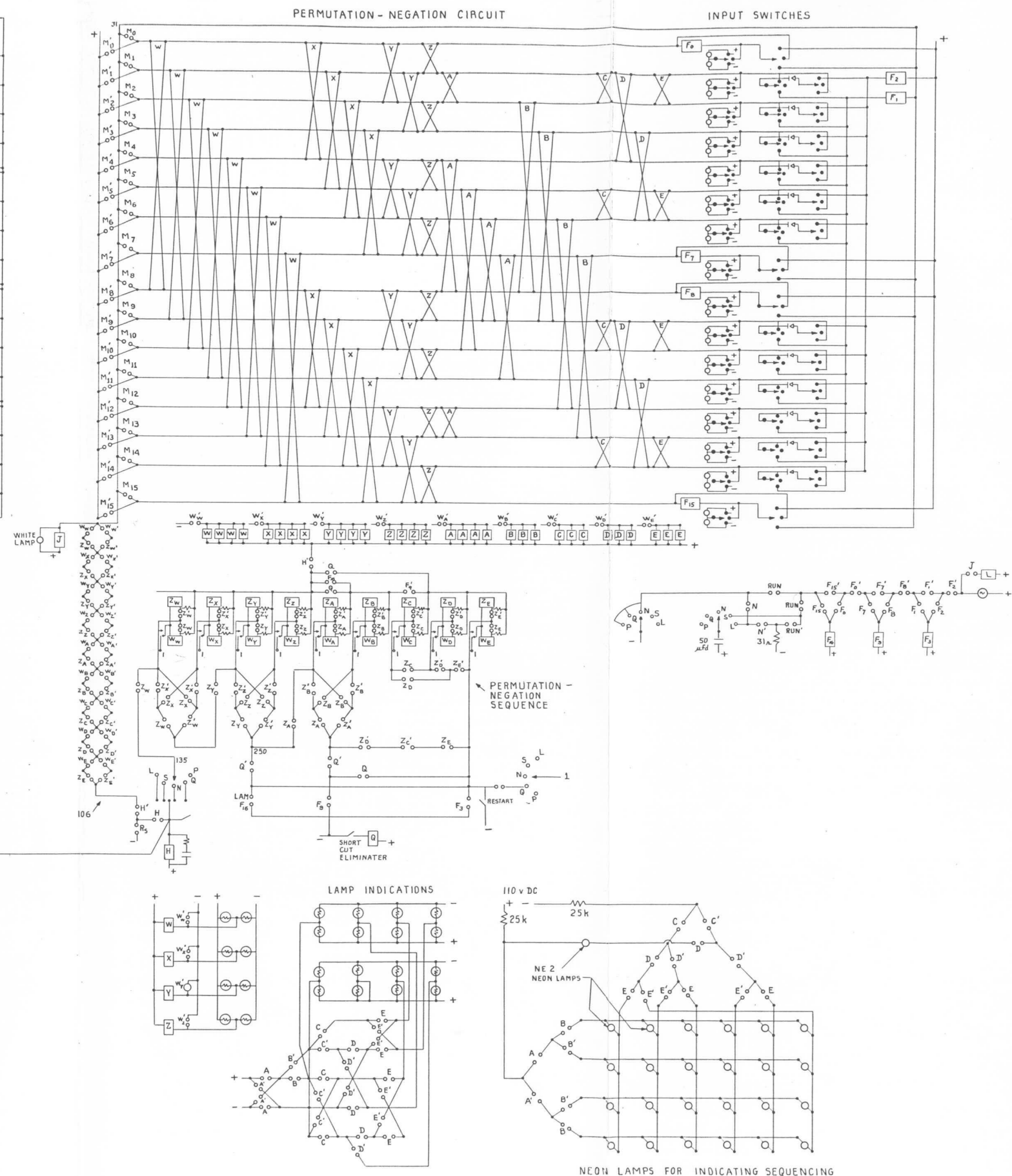
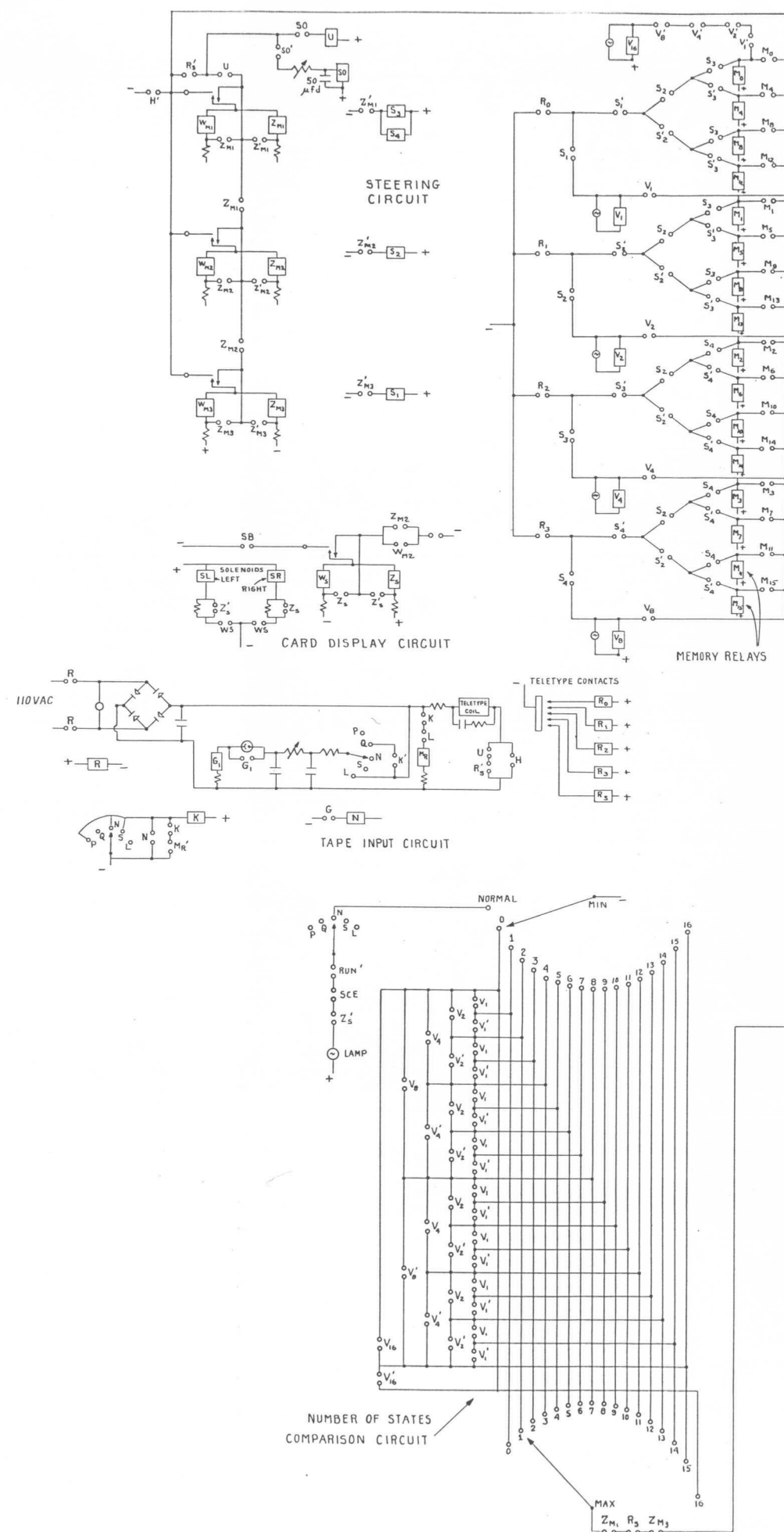
*See "Inequalities," Hardy, Littlewood and Polya, Cambridge 1934, p. 80.

$$\begin{aligned}
 \sum_{jk} R_{jk} \Delta P_j \Delta P_k &= -\sum_i \sum_{jk} \frac{1}{Q_i} p_j(i) p_k(i) \Delta P_j \Delta P_k \\
 &= -\sum_i \frac{1}{Q_i} \left(\sum_j p_j(i) \Delta P_j \right) \left(\sum_k p_k(i) \Delta P_k \right) \\
 &= -\sum_i \frac{1}{Q_i} \Delta Q_i \Delta Q_i \\
 &= -\sum_i \frac{\Delta Q_i^2}{Q_i} .
 \end{aligned} \tag{1}$$

This displays the sum as necessarily non-positive, since all terms are non-positive, and consequently shows that R_{jk} is negative semi-definite and R a concave function. The simplicity of the formula (1) for the second derivative of R in an arbitrary direction is quite striking.

A corollary to this result is the following: Consider the set s of points (P_1, P_2, \dots, P_n) with $\sum P_i = 1$ for which R has its maximum value. Normally, of course, there is only one point in the set, but in other cases it is not so limited. Our theorem allows us to deduce that s is always a convex set of points, for if R is maximized at (P_1, \dots, P_n) and also at (P'_1, \dots, P'_n) , it must clearly have the same value at $(\alpha P_1 + (1-\alpha)P'_1, \dots, \alpha P_n + (1-\alpha)P'_n)$.

C. E. SHANNON



CIRCUIT OF SYNTHESIZER

Fig. 2

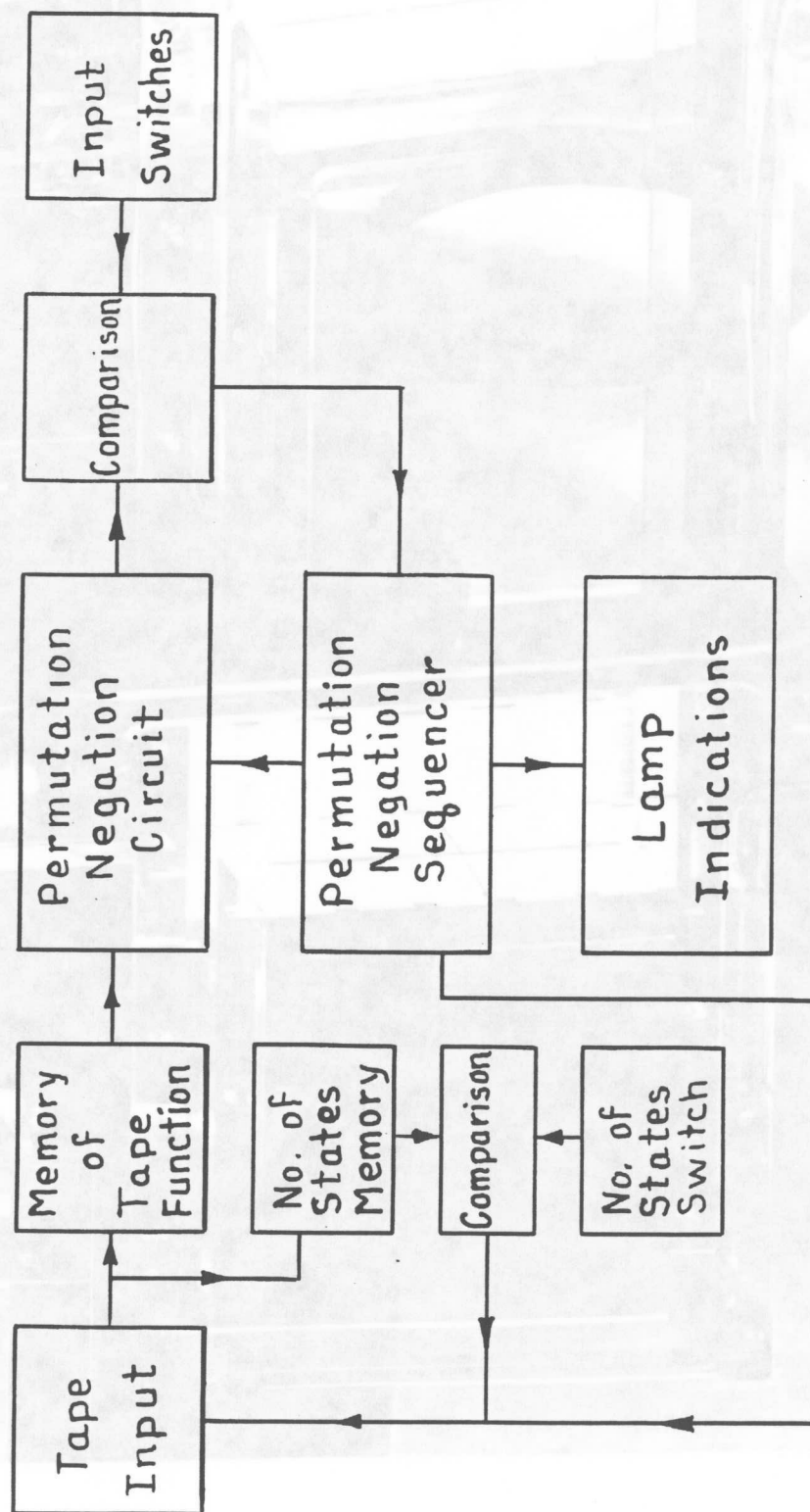
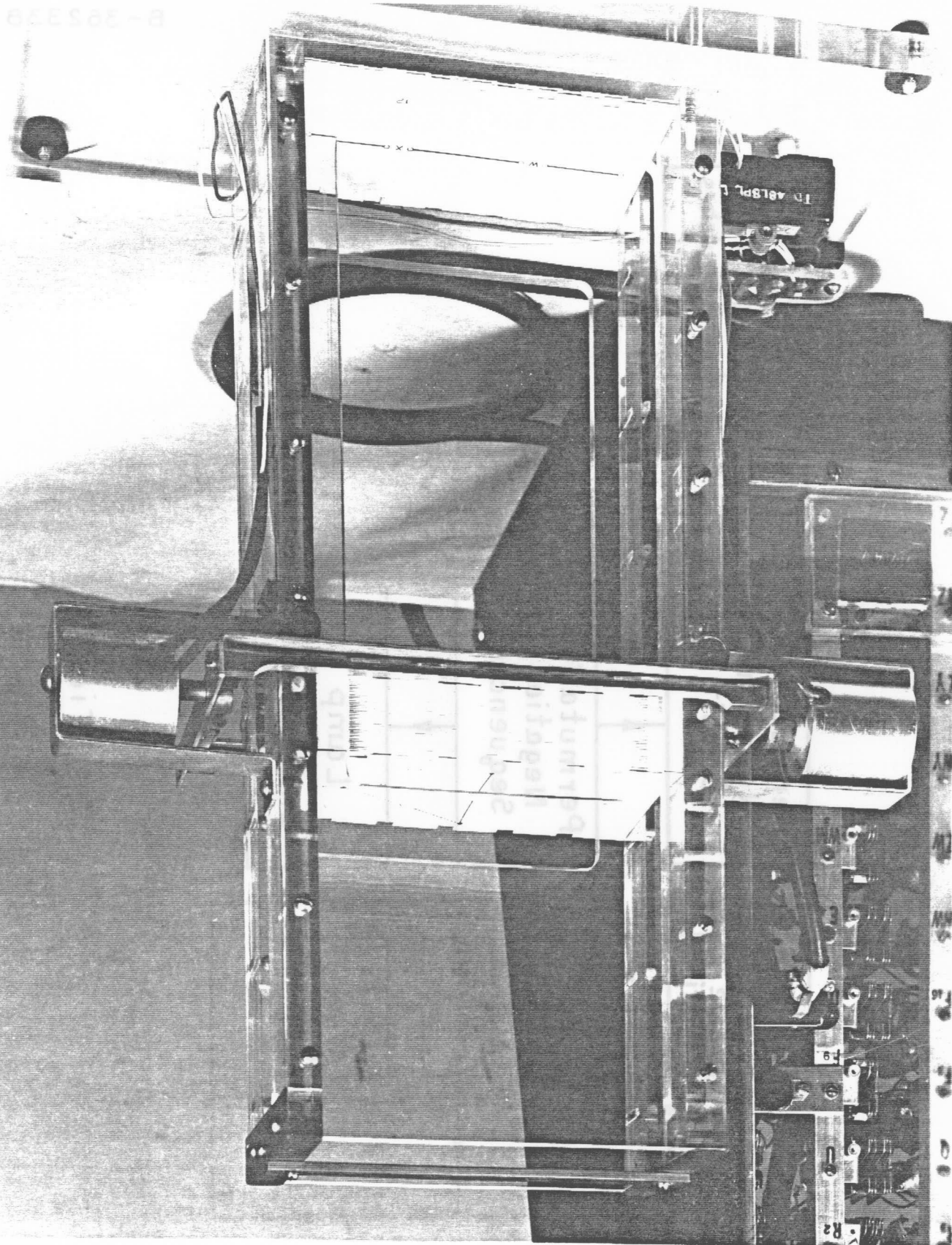
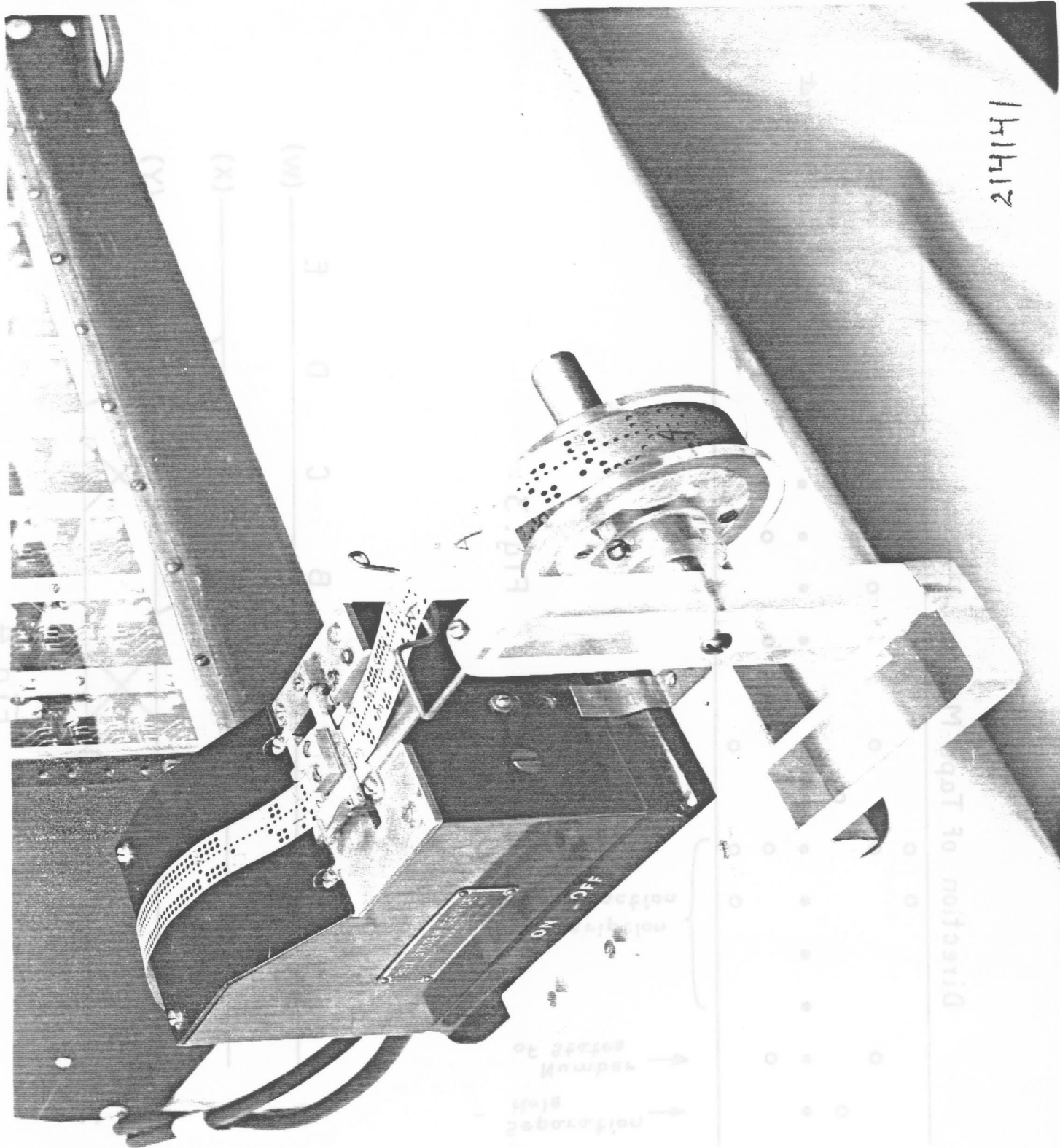


Fig. 1



6-305339



214141



Fig. 3

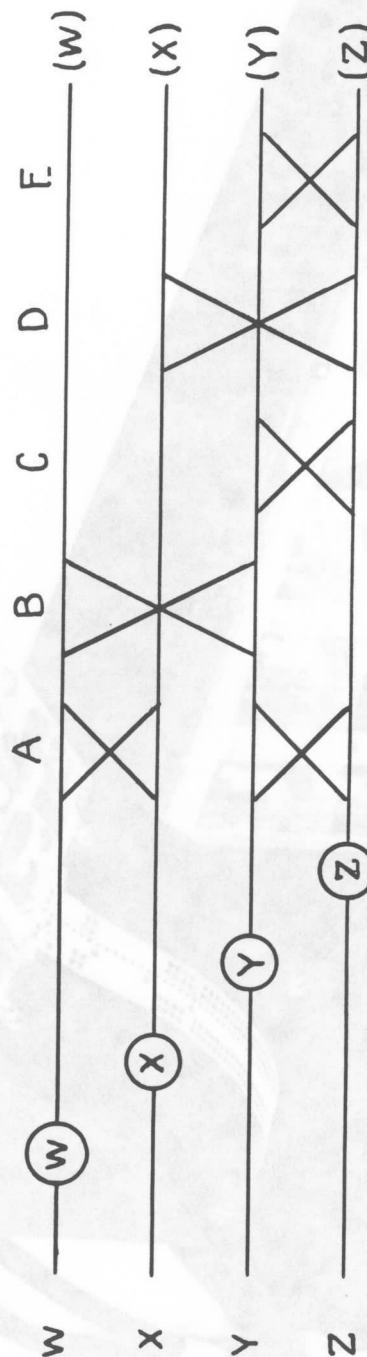


Fig. 4

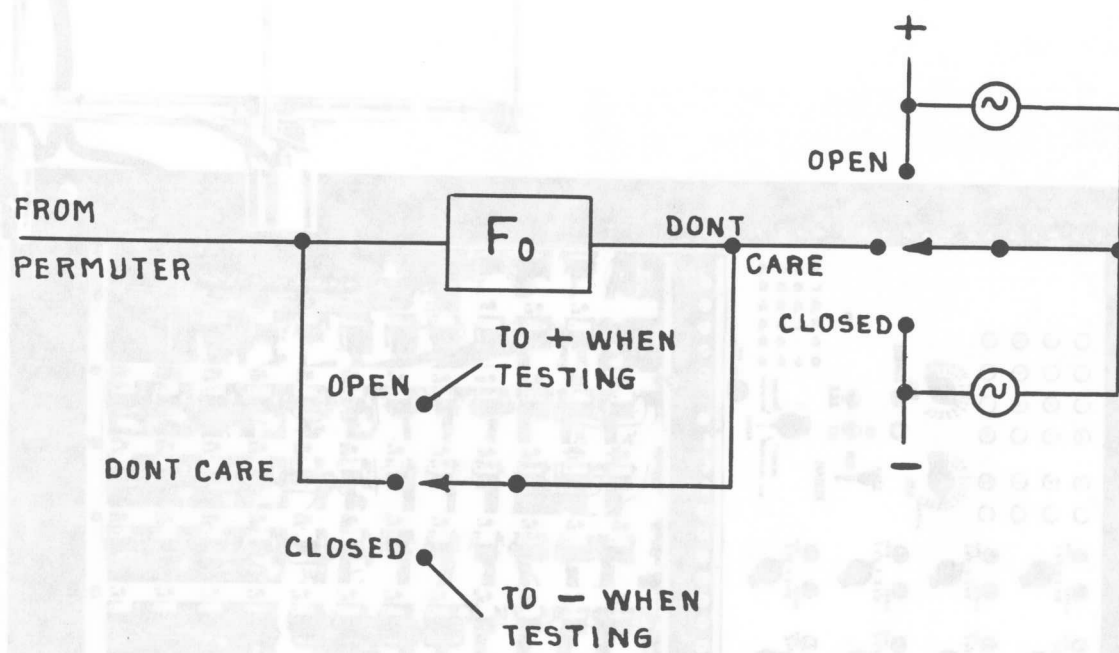
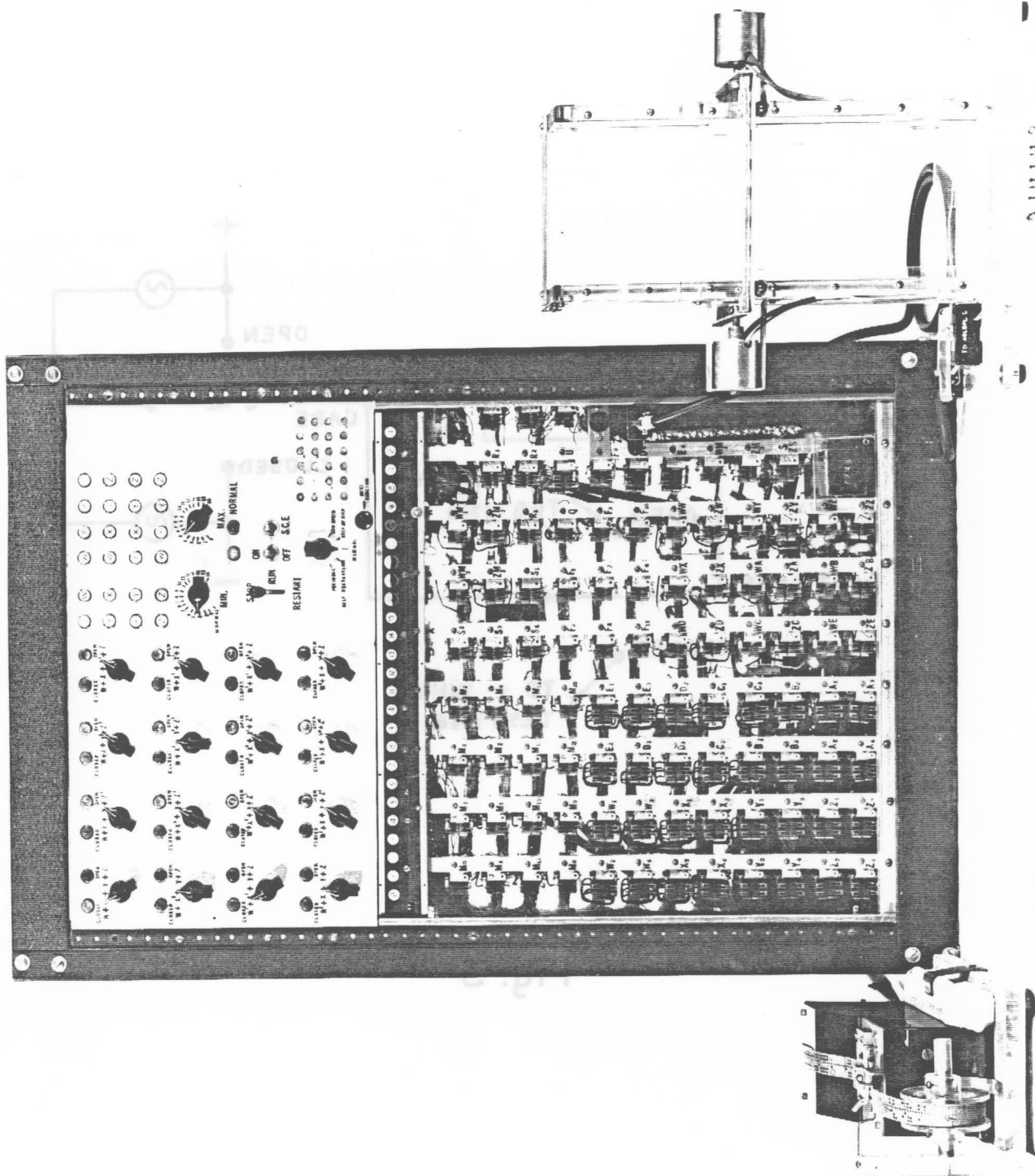
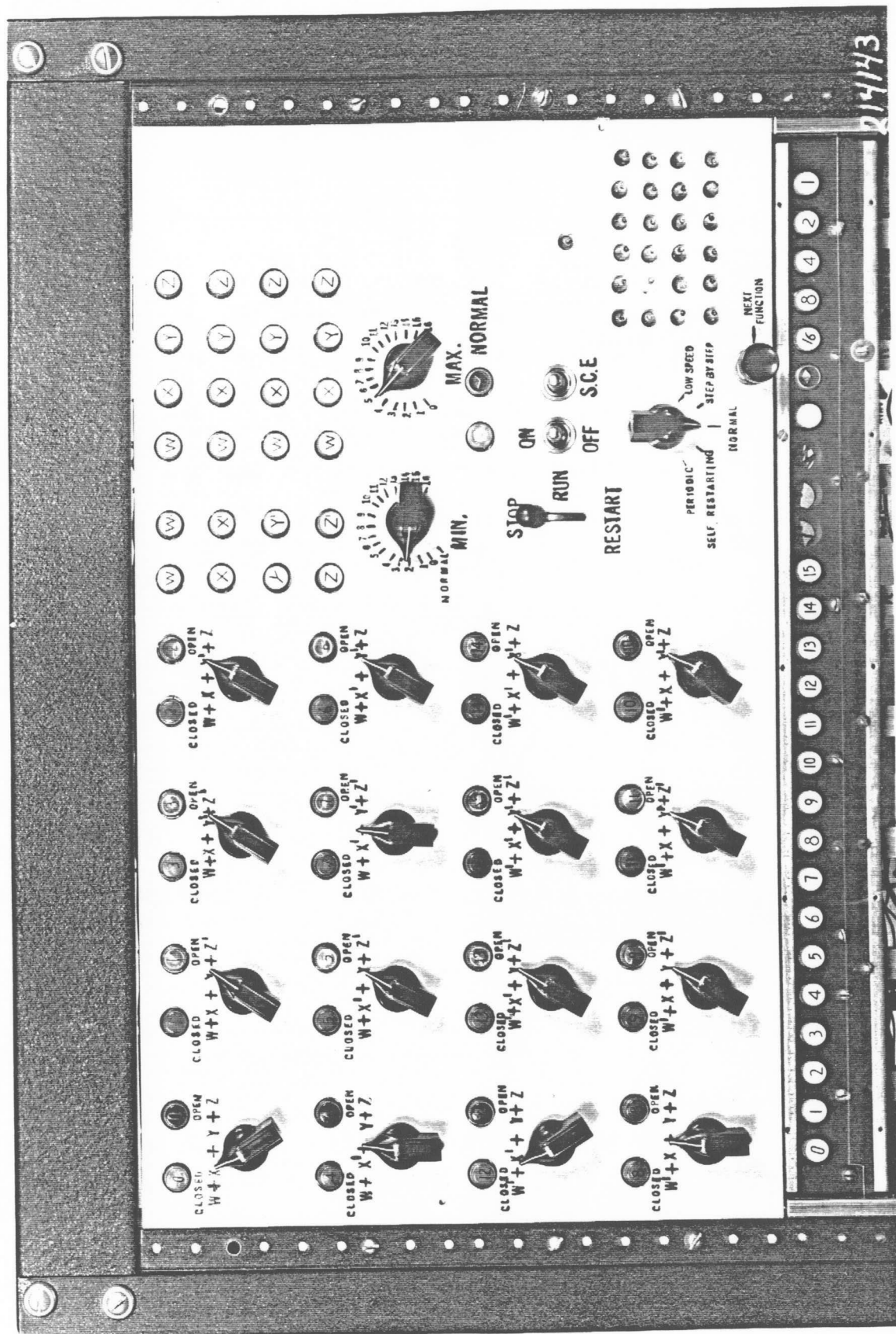


Fig. 5



214142



A SKELETON KEY TO THE INFORMATION SEMINAR - NOTES

The material in these notes has not for the most part been published and is for personal use only. The notes are not complete. Several key sections are not yet available, consequently there are a number of forward and backward references which are quite meaningless. The remaining sections will be handed out as soon as available.

The parts of the notes now available are not arranged in the correct order for easiest reading. The following rearrangement of sections should be made:

<u>Some Useful Inequalities for Distribution Functions</u>	- p. 1a - 3a ✓
<u>A Lower Bound on the Tail of a Distribution</u>	- p. 1y - 9y ✓
<u>A Combination Theorem</u>	p. 1m ✓
<u>Some Results on Determinants</u>	p. 1b - 3b ✓
<u>Upper and Lower Bounds for Powers of a Matrix with Non-negative Elements</u>	
	p. 1w - 3w ✓
<u>The Number of Sequences of a Given Length</u>	p. 1p - 3p ✓
<u>Characteristic for a Language with Independent Letters</u>	p. 1c - 4c ✓
<u>The Probability of Error in Optimal Codes</u>	p. 1 - 4 ✓
Page with figures 1, 2 and 3	
<u>Zero Error Codes and the Zero Error Capacity C_0</u>	p. 1g - 6g ✓
<u>Theorem</u>	p. 1h - 3h ✓
Figure 4	
<u>Lower Bound for P_e for a Completely Connected Channel with Feedback</u>	p. 2r - 3r ✓
<u>A Lower Bound for P_e</u>	p. 1k - 5k ✓
<u>Application of "Sphere-packing" Bounds to Feedback Case</u>	- p. 1p - 3p ✓
<u>Theorem</u>	p. 1q - 4q ✓
<u>Theorem</u>	p. 1j ✓
<u>A Result for the Memoryless Feedback Channel</u>	p. 1r ✓
<u>Continuity of $P_{e \text{ opt}}$ as a function of transition probabilities</u>	- p. 1e ✓
<u>Codes of a fixed composition</u>	p. 1f ✓
<u>Relation of P_e to p</u>	. 1i - 2i ✓
<u>Bound on P_e for Random Code by Simple Threshold Argument</u>	- s1 - s4 ✓
<u>A bound on P_e for a random code</u>	p. 1d - 3d ✓

<u>The Feinstein Bound</u>	pages 11 & 21 ✓
<u>Relations Between Reliability and Minimum Word Separation</u>	- p. 12, 22, 62 & 72 ✓
<u>Inequalities for Decodable Codes</u>	p. 1n - 3n ✓
<u>Convexity of Channel Capacity as a Function of Transition Probabilities</u>	p. 1o ✓
<u>A Geometric Interpretation of Channel Capacity</u>	p. 1x - 6x ✓
<u>Log Moment Generatin Function for the Square of a Guassian Variate</u>	p. § 1 - §2 ✓
<u>Upper Bound on P_e for Gaussian Channel by Expurgated Random Code</u>	p. §1 - §2 ✓
<u>Lower Bound on P_e in Gaussian Channel by Minimum Distance Argument</u>	p. α1 - α2
<u>The Sphere Packing Bound for the Gaussian Power Limited Channel</u>	p. ε 1 - ε 5
<u>The T-terminal Channel</u>	p. §1 - §7
<u>Conditions for Constant Mutual Information</u>	p. 1066
<u>Simple Proof</u>	p. 1024

The following errata have been found:

p. 1y line 10 $\geq 1 - \frac{1}{\alpha}$
 line 11 for any positive α
 line 14 $\geq (1 - \frac{1}{\alpha}) e^n \dots$

p. 2y line 8 $V, < V_2 < \dots V_t$

p. 3w - lines 1, 2, 4, 7, 8, 9, 13, 17 subscripts on β should be in line.

p. 2c - line 7 $\frac{1}{n} \log \text{Prob}$

4c - Eq. (7) $E(s) = -\sum q_1(s) \log \frac{p_1}{q_1(s)} = -(\mu - s\mu')$

Eq. (8) $R(s) = \sum q_1(s) \log q_1(s)^{s-1} = \mu - (s-1)\mu'$

line 6 $\frac{dE}{ds} / \frac{dR}{ds} = -\frac{\mu' + s\mu'' + \mu'}{\mu' + (1-s)\mu'' - \mu'} = -\frac{s}{1-s}$

line 2 $E(1) = \frac{1}{d} \log p_1^{-1} + \log d$

[104]

Bounds on the Tails of Martingales and Related Questions

Claude E. Shannon

Department of Electrical Engineering

Department of Mathematics

and

Research Laboratory of Electronics

Massachusetts Institute of Technology

Cambridge, Massachusetts

This paper is concerned with the problem of overbounding the probability that the sum of n dependent random variables exceeds a certain quantity. Certain restrictions are assumed concerning the distribution of the i th random variable x_i conditional on the preceding random variables. As an example, we might have a gambler playing some "system" in which x_i is his winning on the i th bet. Suppose he can choose any distribution he desires for x_i conditional on the preceding plays, $P(x_i | x_{i-1}, x_{i-2}, \dots, x_1)$, subject however to the conditions 1) it is a "fair" bet, $E(x_i | x_{i-1}, x_{i-2}, \dots, x_1) = 0$; 2) there is a "house limit" on possible wins or losses for one bet, $P(x_i | x_{i-1}, x_{i-2}, \dots, x_1) = 0$ for $x_i < -L$ and $P(x_i | x_{i-1}, x_{i-2}, \dots, x_1) = 1$ for $x_i > W$. It is desired to find an upper bound on the probability that the gambler's winnings will exceed a certain limit λ after n bets. This bound will of course be a function of L , W , n and λ but is to be independent of the system used.

Thought of another way, we can imagine the gambler mapping out a strategy, subject to the house rules, to try to maximize the probability

[101]

of ending up after n bets with a total winning of λ or more. If this is his object, he would clearly be wise, for example, if he ever reached the level λ to not risk any future loss. This he could do by choosing a distribution function thereafter which is 0 for negative x and 1 for positive x .

We will find a bound for this problem and various other similar problems with different side constraints on the allowed distribution functions. The results have applications in various problems related to random walks, gambler's ruin problems and certain coding problems in information theory.

In the example above, the gambler's total capital forms a martingale because of the "fair bet" condition. Bounds on the tails of martingales are known in terms of the variances of the successive amounts won. The bounds we obtain are in terms of conditional moment generating functions. As such, they require more in the way of restrictions on the distributions (for the moment generating functions to exist), but give tighter bounds. Our bounds bear the same relation to the variance type bounds for martingales that the Chernoff bound does to the Chebycheff bound for sums of independent random variables.

The Main Inequality

The method we use is based on a bound for the tail of a distribution due to Chernoff⁽¹⁾. Let $P(x)$ be the distribution function of a random variable and suppose the moment generating function $\nu(s) = \int_{-\infty}^{\infty} e^{sx} dP(x)$ exists over some s interval including the origin in its interior. This

will certainly be true, for example, if $P(x) < e^{ax}$ for some $a > 0$ and sufficiently large negative x , and $1 - P(x) < e^{-bx}$ for some positive b and sufficiently large positive x .

We first derive a somewhat generalized formulation of the Chernoff bound. Let $\mu(s) = \log v(s)$ be the semi-invariant generating function.

Lemma 1: Suppose the semi-invariant generating function $\mu(s)$, for a random variable x , exists for $a \leq s \leq b$ and does not exceed another differentiable function of s , $\mu_0(s)$. Thus $\mu(s) \leq \mu_0(s)$. Then

$$\Pr[x \geq \mu_0'(s)] \leq e^{\mu_0(s) - s\mu_0'(s)} \quad b \geq s \geq 0$$

$$\Pr[x \leq \mu_0'(s)] \leq e^{\mu_0(s) - s\mu_0'(s)} \quad -a \leq s \leq 0$$

This result is like the Chernoff bound except for replacement of $\mu(s)$ by an upper bounding function $\mu_0(s)$, and may be proved by similar means. Thus by the generalized Chebycheff inequality

$$\begin{aligned} e^{s\lambda} \Pr[x \geq \lambda] &\leq \int_{\lambda}^{\infty} e^{sx} dP(x) \quad s \geq 0 \\ &\leq \int_{-\infty}^{\infty} e^{sx} dP(x) = v(s) = e^{\mu(s)} \\ &\leq e^{\mu_0(s)} \end{aligned}$$

This is true for any λ . Set $\lambda = \mu_0'(s)$. Then

$$\Pr[x \geq \mu_0'(s)] \leq \frac{e^{\mu_0(s)}}{e^{s\mu_0'(s)}} = e^{\mu_0(s) - s\mu_0'(s)} \quad b \geq s \geq 0$$

A similar argument gives the dual inequality for negative s .

We now develop a formula for the moment generating function of the sum of a set of dependent random variables $u = x_1 + x_2 + \dots + x_n$, where the distribution function of x_1, \dots, x_n is given by

$$P(z_1, z_2, \dots, z_n) = \Pr[x_1 \leq z_1, x_2 \leq z_2, \dots, x_n \leq z_n]$$

It is as assumed that for this multivariate distribution the moment generating functions for various random variables conditional on others exist. To avoid notational complexity we carry out the argument only for $n = 3$, using x , y and z for the three random variables, but the method is clearly general. If $v(s)$ is the moment generating function for the sum variable $u = x + y + z$, then (all integrals are from $-\infty$ to ∞);

$$\begin{aligned} v(s) &= \int \int \int e^{s(x+y+z)} dP(x, y, z) \\ &= \int \int \int e^{s(x+y+z)} dP(x) dP(y|z) dP(z|x, y) \\ &= \int e^{sx} dP(x) \int e^{sy} dP(y|x) \int e^{sz} dP(z|x, y) \end{aligned}$$

The innermost integral is the moment generating function for z conditional on x and y , and may be denoted by $v_3(s|x, y)$ (the 3 referring to the third variable, z). Thus

$$v(s) = \int e^{sx} dP(x) \int e^{sy} dP(y|x) v_3(s|x, y)$$

Suppose now that we have a bounding function for $v_3(s|x, y)$, say $\gamma_3(s)$,

independent of x and y .

$$\nu_3(s|x, y) \leq \gamma_3(s)$$

Then the innermost integral may be bounded by $\gamma_3(s)$ and this term taken out of the integration. (ν_3 is clearly non-negative, being an expectation of e^{sz} .) Thus

$$\nu(s) \leq \gamma_3(s) \int e^{sx} dP(x) \int e^{sy} dP(y|x)$$

Similarly, suppose the moment generating function of y conditional on x is bounded by $\gamma_2(s)$

$$\nu_2(s|x) = \int e^{sy} dF(y|x) \leq \gamma_2(s)$$

and the moment generating function of x is bounded by $\gamma_1(s)$

$$\nu_1(s) = \int e^{sx} dP(x) \leq \gamma_1(s)$$

Then these may also be used to bound the integrals, giving

$$\nu(s) \leq \gamma_1(s) \gamma_2(s) \gamma_3(s)$$

Taking logarithms, the semi-invariant generating function $\mu(s)$ for the sum variable u is therefore bounded by the sum of the logarithms of the $\gamma(s)$ functions, that is, by uniform bounds on the conditional semi-invariant functions for the different variables

$$\mu(s) \leq \mu_1(s) + \mu_2(s) + \mu_3(s)$$

The same argument carries through for the sum of any number of random variables and may be summarized as follows.

Lemma 2: The semi-invariant generating function $\mu(s)$ for the sum of n random variables is bounded by

$$\mu(s) \leq \sum \mu_i(s)$$

where $\mu_i(s)$ is a uniform bound on the semi-invariant function for the i th random variable conditional on the first $i-1$:

$$\log \int e^{sx_i} dP(x_i | x_1, x_2, \dots, x_{i-1}) \leq \mu_i(s).$$

In most applications the same bound, say $\mu_0(s)$, will apply to all the random variables. In this case $\mu(s) \leq n\mu_0(s)$. Combining Lemmas 1 and 2 we obtain our first main result, a bound on the tail probability of a sum of dependent random variables provided the conditional moment generating functions exist.

Theorem 1: If u is the sum of n dependent random variables $x_i (i=1, 2, \dots, n)$ whose semi-invariant generating functions conditional on preceding variables $\mu_i(s | x_1, \dots, x_{i-1})$ exist and are bounded by differentiable functions $\mu_i(s)$, $(i=1, 2, \dots, n)$ then

$$\Pr[u \geq \sum \mu_i^0(s)] \leq e^{\sum \mu_i(s) - s \sum \mu_i^0(s)} \quad s \geq 0$$

$$\Pr[u \leq \sum \mu_i^0(s)] \leq e^{\sum \mu_i(s) - s \sum \mu_i^0(s)} \quad s \leq 0$$

Applications

In applications of this result we would generally attempt to find the smallest bounding functions $\mu_i(s)$ in order to obtain the tightest bound on the tail probability. As a first example consider a gambler allowed to choose a wager with an arbitrary distribution function $\phi(x)$ (the probability of gaining x or less), subject however to the following conditions:

- 1) The expected gain is zero. $\int x d\phi(x) = 0$
- 2) $\phi(x) \leq \phi_1(x)$ where $\phi_1(x)$ is a distribution function with negative mean for which $\int e^{sx} d\phi_1(x)$ exists for some negative s .
- 3) $\phi(x) \geq \phi_2(x)$ where $\phi_2(x)$ is a distribution function with positive mean for which $\int e^{sx} d\phi_2(x)$ exists for some positive s and $\phi_2(x) \leq \phi_1(x)$.

Thus our gambler is allowed to choose a distribution function $\phi(x)$ at each wager lying between two given curves $\phi_1(x)$ and $\phi_2(x)$, (as suggested by Fig. 1)

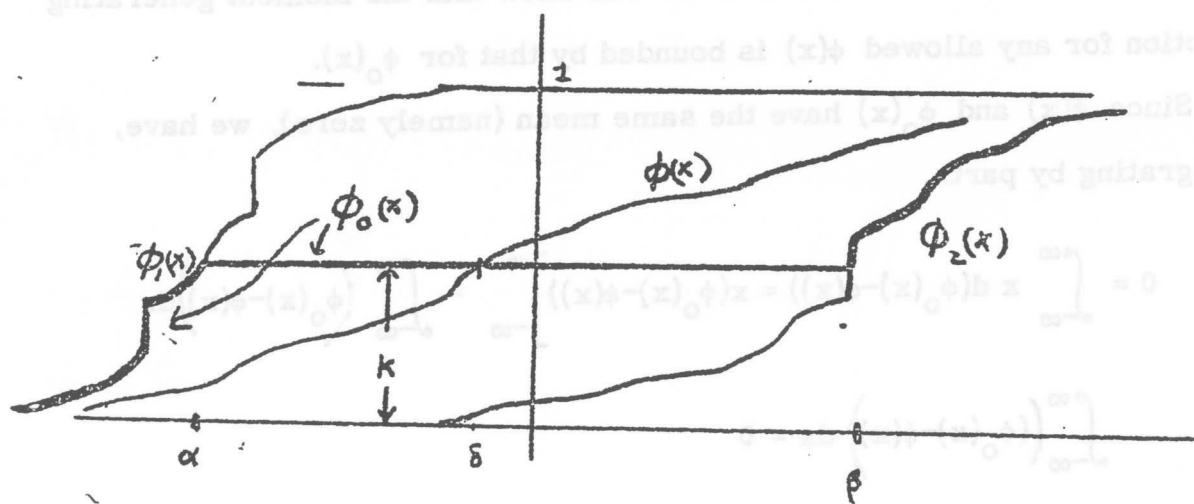


Fig. 1.

which approach 0 and 1 with a certain rapidity. He is also constrained

to choose a distribution function with zero mean. The situation described earlier involving house limits is a case of this type where the distributions ϕ_1 and ϕ_2 are step functions at L and W , the maximum allowed loss or win per wager.

To apply the theorem we need a function which bounds the moment generating functions which he can achieve with these restrictions. Consider the distribution function $\phi_0(x)$ defined as follows:

$$\phi_0(x) = \phi_1(x) \quad x < a$$

$$\phi_0(x) = k \quad a \leq x \leq \beta$$

$$\phi_0(x) = \phi_2(x) \quad x > \beta$$

where a is the first point at which $\phi_1(x)$ reaches the value k and β is the first point at which $\phi_2(x)$ reaches k . $\phi(x)$ is a distribution function, and by adjusting k we can clearly make the mean of the distribution $\phi(x)$ equal zero. With this value of k we will show that the moment generating function for any allowed $\phi(x)$ is bounded by that for $\phi_0(x)$.

Since $\phi(x)$ and $\phi_0(x)$ have the same mean (namely zero), we have, integrating by parts,

$$0 = \int_{-\infty}^{\infty} x d(\phi_0(x) - \phi(x)) = x(\phi_0(x) - \phi(x)) \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} (\phi_0(x) - \phi(x)) dx$$

$$\int_{-\infty}^{\infty} (\phi_0(x) - \phi(x)) dx = 0$$

where we use the exponential approach of ϕ and ϕ_0 to 0 and 1 as x goes to $-\infty$ and $+\infty$ to insure the vanishing of the term $x(\phi_0(x) - \phi(x))$ at these limits.

Now consider the quantity (again using integration by parts)

$$\begin{aligned} \int_{-\infty}^{\infty} e^{sx} d(\phi_0(x) - \phi(x)) &= e^{sx}(\phi_0(x) - \phi(x)) \Big|_{-\infty}^{\infty} - s \int_{-\infty}^{\infty} (\phi_0(x) - \phi(x)) e^{sx} dx \\ &= -s \int_{-\infty}^{\delta} e^{sx} [\phi_0(x) - \phi(x)] dx - s \int_{\delta}^{\infty} e^{sx} [\phi_0(x) - \phi(x)] dx \end{aligned}$$

$$-a \leq s \leq b$$

where a and b are the limits of convergence of the moment generating functions and δ is the first point at which $\phi(x)$ reaches the value k , the horizontal segment of $\phi_0(x)$. Notice that, for $x \leq \delta$, $\phi_0(x) - \phi(x)$ is positive (or zero) while for $x > \delta$, $\phi_0(x) - \phi(x)$ is negative (or zero). The first term, $-s \int_{-\infty}^{\delta} e^{sx} [\phi_0(x) - \phi(x)] dx$ is greater than or equal to $-s \int_{-\infty}^{\delta} e^{s\delta} [\phi_0(x) - \phi(x)] dx$, since, when s is positive, $e^{s\delta} \geq e^{sx}$ for $-\infty < x \leq \delta$, $\phi_0 - \phi$ is positive and the coefficient $-s$ is negative. If s is negative, e^{sx} has its least value at $x = \delta$ but the coefficient $-s$ is now positive. In a similar way, the second integral $-s \int_{\delta}^{\infty} e^{sx} [\phi_0(x) - \phi(x)] dx$ is greater than or equal to $-s \int_{\delta}^{\infty} e^{\delta x} [\phi_0(x) - \phi(x)] dx$ as one verifies by examination of the two cases $s \geq 0$ and $s < 0$, remembering that $\phi_0(x) - \phi(x)$ is negative or zero in this range. Thus we conclude

$$\begin{aligned} \int_{-\infty}^{\infty} e^{sx} d[\phi_0(x) - \phi(x)] &\geq -s \int_{-\infty}^{\delta} e^{s\delta} [\phi_0(x) - \phi(x)] dx - s \int_{\delta}^{\infty} e^{\delta x} [\phi_0(x) - \phi(x)] dx \\ &= -s e^{\delta s} \int_{-\infty}^{\infty} [\phi_0(x) - \phi(x)] dx \\ &= 0 \end{aligned}$$

$$\int_{-\infty}^{\infty} e^{sx} d\phi_0(x) \geq \int_{-\infty}^{\infty} e^{sx} d\phi(x)$$

In other words, the moment generating function for the distribution $\phi_0(x)$ dominates that of any other distribution with the same mean as ϕ_0 and bounded by the ϕ_1 and ϕ_2 curves. Therefore the moment generating function for ϕ_0 may be used in our bounds for the tail of a sum distribution if the individual conditional distributions satisfy this type of restriction.

Using this bound on the conditional moment generating functions in Theorem 1 our solution may be summarized as follows. Suppose at each play of a game the distribution functions available to a gambler all have zero mean and lie between two functions $\phi_1(x)$ and $\phi_2(x)$. Let $\phi_0(x)$ be the zero mean function consisting of ϕ_1 followed by a flat segment, followed by ϕ_2 . Let

$$\mu(s) = \log \int_{-\infty}^{\infty} e^{sx} d\phi_0(x)$$

Then the probability of his winnings after n wagers exceeding $n\mu^1(s)$ is bounded by

$$\Pr[u \geq n\mu^1(s)] \leq e^{n[\mu(s) - s\mu^1(s)]} \quad s \geq 0$$

This same bound applies, of course, also with a semi-martingale condition, that is, if the gambler's expectation is only required to be non-positive.

If $\phi_1(0) = 1$ and $\phi_2(0) = 0$ (so the gambler can play a wager that amounts to stopping the game, that is, a distribution which is a unit step at zero), then this same bound applies to the probability of exceeding $n\mu^1(s)$ on any of the first n trials. This is because the bound covers all strategies.

Any particular strategy could be modified so that if the gambler reaches the level $n\mu'(s)$ at any time before the n th trial he then effectively holds his winnings by playing the distribution with unit step at zero. The bound must exceed the probability of exceeding the level $n\mu'(s)$ for this strategy at the n th step but this is a bound on the probability of ever exceeding the level in the first n steps. This device can be used in many applications of the method we are describing, provided only that the unit step at zero is an allowed distribution function.

The bound given, while certainly not the best possible for all values of the parameters, is, however, best possible in the coefficient of n in the exponent. That is, the result would be false if $\mu(s) - s\mu'(s)$ in the right hand exponential term were replaced by $\mu(s) - s\mu'(s) - \epsilon$ for any positive ϵ . This may be seen as follows. The gambler could, within the rules, choose the distribution $\phi_0(x)$ at each wager. If he does so, then we have a sum of n independent random variables, each with semi-invariant generating function $\mu(s)$. Lower bounds on the tail of this sum distribution are known to exceed $e^{n[\mu(s)-s\mu'(s)-\epsilon]}$ when n is sufficiently large.⁽¹⁾

The Case with House Limits on Win or Loss for each Wager

For the case of the gambler who can choose an arbitrary distribution with zero mean and house limits on wins and losses W and L ($L < 0$) respectively, the distribution to maximize $\mu(s)$ is, from the above analysis, a binomial distribution with jumps at the ends of the interval W and L adjusted to give a zero mean. The two probabilities are $\frac{W}{W-L}$ at L and

$\frac{-L}{W-L}$ at W .

To gain a little in generality and simplify notation, consider a binomial with probability p at values L and probability $q = (1-p)$ at W . The semi-invariant generating function is

$$\mu(s) = \log (pe^{sL} + qe^{sW})$$

$$\mu'(s) = \frac{pLe^{sL} + qWe^{sW}}{pe^{sL} + qe^{sW}}$$

The expression for the bound on the tail may be simplified by a change of variables eliminating s . Let

$$\lambda = \frac{pe^{sL}}{pe^{sL} + qe^{sW}}$$

$$\eta = 1 - \lambda = \frac{qe^{sW}}{pe^{sL} + qe^{sW}}$$

Then

$$\frac{\lambda}{\eta} = \frac{p}{q} e^{s(L-W)}$$

$$s = \frac{1}{L-W} \log \frac{\lambda q}{p \eta}$$

$$\mu'(s) = \lambda L + \eta W$$

$$\mu - s\mu'(s) = \log(pe^{sL+qW}) - s(\lambda L + \eta W)$$

$$= \log(pe^{sL+qW}) - \frac{(\lambda L + \eta W)}{L - W} \log \frac{\lambda q}{p\eta}$$

$$= \log \frac{p}{\lambda} + \frac{L}{L - W} \log \frac{\lambda q}{p\eta} - \frac{(\lambda L + \eta W)}{L - W} \log \frac{\lambda q}{p\eta}$$

$$= \lambda \log \frac{p}{\lambda} + \eta \log \frac{q}{\eta}$$

Letting p equal $\frac{W}{W-L}$ and q equal $\frac{-L}{W-L}$ and using our result bounding the tail of the sum of n random variables, we obtain the following bound for the probability of the gambler exceeding a certain level after n wagers:

$$\Pr[u \geq n(\lambda L + \eta W)] \leq \left\{ \frac{\left(\frac{W}{\lambda}\right)^\lambda \left(\frac{-L}{\eta}\right)^\eta}{W - L} \right\}^n \quad \lambda \geq p; \eta = 1 - \lambda$$

If $L = -W$, that is, the win and loss limits are the same, this formula can be simplified somewhat at the expense of a certain weakening. It then becomes

$$\Pr[u \geq nW(1-2\lambda)] \leq \left[\frac{\lambda^{-\lambda} \eta^{-\eta}}{2} \right]^n$$

Let $\lambda = \frac{1}{2}(1+\theta)$, $\eta = \frac{1}{2}(1-\theta)$.

Then

$$\Pr[u \geq nW\theta] \leq [(1+\theta)^{-(1+\theta)}(1-\theta)^{-(1-\theta)}]^{n/2}$$

$$= e^{-\frac{n}{2}[(1+\theta) \ln(1+\theta) + (1-\theta) \ln(1-\theta)]}$$

Consider the bracketed term in the exponent and expand the logarithms as series.

$$\begin{aligned}
 [(1+\theta) \ln(1+\theta) + (1-\theta) \ln(1-\theta)] &= (1+\theta) \left(\theta - \frac{\theta^2}{2} + \frac{\theta^3}{3} - \frac{\theta^4}{4} + \dots \right) \\
 &\quad + (1-\theta) \left(-\theta - \frac{\theta^2}{2} - \frac{\theta^3}{3} - \frac{\theta^4}{4} - \dots \right) \\
 &= 2 \left(-\frac{\theta^2}{2} - \frac{\theta^4}{4} - \frac{\theta^6}{6} - \dots \right) \\
 &\quad + 2 \left(\theta^2 + \frac{\theta^4}{3} + \frac{\theta^6}{5} + \dots \right) \\
 &= \theta^2 + \frac{\theta^4}{6} + \frac{\theta^6}{15} + \dots + \frac{\theta^{2n}}{n(2n-1)} + \dots \\
 &\geq \theta^2
 \end{aligned}$$

Hence

$$\Pr[u \geq nW\theta] \leq e^{-\frac{n\theta^2}{2}} \quad \theta \geq 0$$

It may be noted that this bound is similar to the exponential part of the normal approximation to the sum of n binomial samples, probabilities $\frac{1}{2}$ at $\frac{1}{2}W$, without, however, the coefficient term that would ordinarily appear. This might suggest that the gambler's best strategy to maximize the probability of exceeding $nW\theta$ would be to continually play the extreme binomial distribution, or at least until he was within W of it, and then switch to a binomial which would just carry him over the limit if he won. While this appears to be a rather good strategy, it is not quite optimal

as a study of small n values reveals. Determining the optimal strategy appears to involve considerable combinatorial complexity.

The Probability of ever exceeding a Limit with a Negative Expectation

Suppose now that the conditional expectation of all wagers is negative and we are interested in a bound on the probability of ever (in an infinite series of wagers) exceeding a certain (positive) value. If the expectation were zero, then by well known results in the gambler's ruin problem the only bound is unity, provided, for example, the gambler can play a binomial distribution. With a negative mean, however, significant bounds can be obtained as follows.

We consider the case again where the allowed distribution functions must lie between two given distribution functions $\phi_1(x)$ and $\phi_2(x)$ but now must have a mean $m < 0$. The maximum $\mu(s)$ is obtained by the same construction using ϕ_1 and ϕ_2 , but with a placement of the horizontal segment to give the mean m .

If $\phi(0)$ is 1, then $\phi_2(0)$ must have been 1, and no allowed bet whatever will ever give a positive return. Thus clearly the probability of ever exceeding any positive bound is zero. We will therefore assume that $\phi(0) < 1$. This assumption also excludes $\phi(x)$ being a unit step, since the step would have to occur at the negative number m , making $\phi(0)$ equal 1.

Under the assumption $\phi(0) < 1$, the $\mu(s)$ curve has the general form shown in Fig. 2.

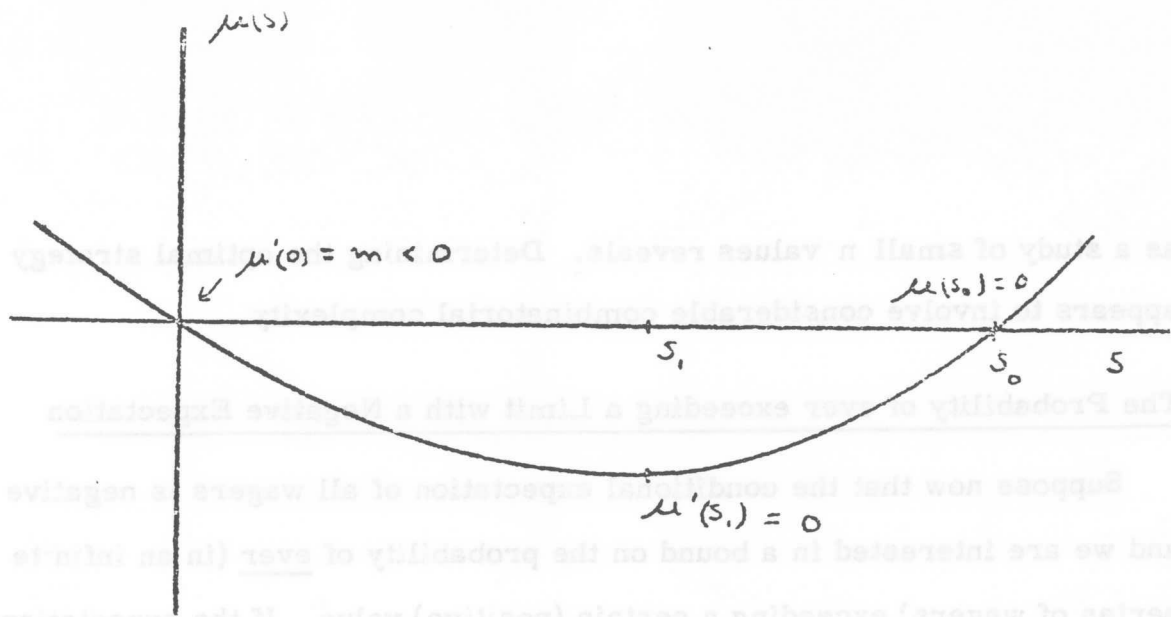


Fig. 2.

The curve is convex downward; it passes through zero at $s = 0$ with a negative slope m ; it has a unique minimum at $s = s_1$ (say); and passes through zero again at $s_0 > s_1$. These facts follow readily from the relations

$$v(s) = \int e^{sx} d\phi(x) \quad \mu(s) = \ln v(s)$$

$$v(0) = \int d\phi(x) = 1 \quad \mu(0) = 0$$

$$v'(s) = \int x e^{sx} d\phi(x) \quad \mu'(s) = \frac{v'(s)}{v(s)}$$

$$v'(0) = \int x d\phi(x) = m \quad \mu'(0) = m$$

$$v''(s) = \int x^2 e^{sx} d\phi(x) \quad \mu''(s) = \frac{v(s) v''(s) - v'(s)^2}{(v(s))^2}$$

The numerator of $\mu''(s)$ is positive by using the Schwartz inequality

(the unit step which would give zero being excluded). Hence the μ curve is strictly convex downward. Also, for sufficiently large positive s , $v(s)$ will exceed 1 and $\mu(s)$ will be positive, since $\phi(0) < 1$. Consequently, the minimum μ at $s = s_1$ and the positive zero crossing at $s = s_0$ both exist.

Suppose we are interested in a bound on the probability of ever reaching or exceeding A with the sums $u_1 = x_1$, $u_2 = x_1 + x_2$, ..., $u_n = \sum x_n$, We have

$$\Pr[\text{any } u_n \geq A] \leq \sum_n \Pr[u_n \geq A]$$

From our above results $\Pr[u_n \geq A] \leq e^{n[\mu(s) - s\mu'(s)]}$ for the s such that $A = n\mu'(s)$. The particular n for which this bound is largest may be obtained by maximizing $n[\mu(s) - s\mu'(s)]$ given $A = n\mu'(s)$, or, in other words, maximizing $A \left[\frac{\mu(s)}{\mu'(s)} - s \right]$. Since $\mu''(s) > 0$ this maximum exists and occurs at a unique s found by differentiation, namely, the s for which $\mu(s) = 0$. This s is the s_0 of Fig. 2, and the corresponding n we call n_0 . Thus s_0 and n_0 satisfy

$$n_0 \mu'(s_0) = A$$

$$\mu(s_0) = 0$$

In general, n_0 will not be an integer, but the bound obtained for evaluation at n_0 and s_0 certainly is greater than that for any integer points. Hence for any particular n ,

$$\Pr[u_n \leq A] \leq e^{n_0(\mu(s_0) - s_0 \mu'(s_0))} = e^{n_0 s_0 \mu'(s_0)}$$

Now consider the s_1 where $\mu'(s_1) = 0$ (Fig. 2) and n_1 defined by

$$n_1 \mu(s_1) = n_0 s_0 \mu'(s_0)$$

Again, in general, n_1 will not be an integer. We let, however, $[n_1]$ denote the largest integer contained in n_1 .

Returning to our inequality on the probability of u_n ever exceeding A we may rewrite as follows

$$\Pr[\text{any } u_n \geq A] \leq \sum_n \Pr[u_n \geq A]$$

$$= \sum_{n=1}^{[n_1]} \Pr[u_n \geq A] + \sum_{[n_1]+1}^{\infty} \Pr[u_n \geq A]$$

$$\leq \sum_{n=1}^{[n_1]} \Pr[u_n \geq A] + \sum_{[n_1]+1}^{\infty} \Pr[u_n \geq 0]$$

$$\leq [n_1] e^{-n_0 s_0 \mu'(s_0)} + \sum_{[n_1]+1}^{\infty} e^{n \mu(s_1)}$$

$$\leq n_1 e^{-n_0 s_0 \mu'(s_0)} + \frac{e^{([n_1]+1)\mu(s_1)}}{1 - e^{\mu(s_1)}}$$

$$\leq n_1 e^{-n_0 s_0 \mu'(s_0)} + \frac{e^{n_1 \mu(s_1)}}{1 - e^{\mu(s_1)}}$$

$$\leq e^{-n_0 s_0 \mu'(s_0)} \left[n_1 + \frac{1}{1 - e^{\mu(s_1)}} \right]$$

$$= e^{-n_1 \mu(s_1)} \left[n_1 + \frac{1}{1 - e^{\mu(s_1)}} \right]$$

$$\Pr[\text{any } u_n \geq A] = e^{-s_0 A} \left[\frac{s_0 A}{\mu(s_1)} + \frac{1}{1 - e^{\mu(s_1)}} \right]$$

This is our desired bound. It is essentially exponentially decreasing in A . In fact more refined analysis can be given to show that the bounded term can be replaced by a more involved expression which does not increase with A .

References

Chernoff, H. (1952). A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sum of Observations. Ann. Math. Stat. 23, 493-507.

Some Useful Inequalities for Distribution Functions

In this section a number of inequalities will be given which are useful in estimating the "tails" of distribution functions or other related statistics.

Binomial Inequalities: let

$$G = \frac{1}{\sqrt{2\pi n \lambda \mu}} \cdot \frac{1}{\lambda^{\lambda n} \mu^{\mu n}} \quad (1)$$

Then

$$G \exp\left(-\frac{1}{12\lambda n} - \frac{1}{12\mu n}\right) \leq \binom{n}{\lambda n} \leq G \quad (2)$$

and

$$\frac{\sqrt{\pi}}{2} G \leq \binom{n}{\lambda n}$$

where $\mu = 1 - \lambda$ and neither λ nor μ is zero. (Note that if either is zero, G is undefined.) Similar inequalities hold for the terms of a binomial distribution, $\binom{n}{\lambda n} p^{\lambda n} q^{\mu n}$, and may be obtained by multiplying the above inequalities by $p^{\lambda n} q^{\mu n}$. They may also be generalized to the multinomial coefficient:

$$G_1 = \frac{1}{\sqrt{2\pi n \prod \lambda_i}} \cdot \frac{1}{\prod \lambda_i^{n \lambda_i}} \quad (3)$$

$$G_1 \exp\left(-\frac{s}{12}\right) \leq G_1 \exp\left(-\sum \frac{1}{12\lambda_i n}\right) \leq \frac{n!}{n(\lambda_i n)!} \leq G_1 \quad (4)$$

where s is the number of components, $\sum \lambda_i = 1$ and none of the λ_i vanishes.

The "tail" of a binomial distribution may be estimated by the following formulas:

$$\sum_{k=\lambda n}^n \binom{n}{k} p^k q^{n-k} \leq \frac{1}{12\lambda n} G p^{\lambda n} q^{\mu n} \text{ provided } \lambda > p + \frac{1}{n} \quad (5) \quad ??$$

$$\sum_{k=\lambda n}^n \binom{n}{k} p^k q^{n-k} \leq \left[\left(\frac{p}{\lambda}\right)^\lambda \left(\frac{q}{\mu}\right)^\mu\right]^n \text{ provided } \lambda > p \quad (6)$$

$$\sum_{k=\lambda n+1}^n \binom{n}{k} p^k q^{n-k} < \frac{\mu p}{\lambda - p} p^{\lambda n} q^{\mu n} G \quad \text{if } \lambda > p$$

$$< \frac{\lambda q}{p - \lambda} p^{\lambda n} q^{\mu n} G \quad \text{if } \lambda > p$$

The first of these gives a closer estimate of the tail but is somewhat more complex. The inequality (6) (Chernoff) is often convenient because of its simplicity. Lower bounds for tails may be taken to be merely lower bounds for the first term as in the lower inequalities of (2) or (4).

We shall now prove the inequalities (1) and (2). The Stirling approximation for $n!$ is as follows:

$$n! = (2\pi)^{1/2} n^{n+1/2} e^{-n} \exp\left(\frac{1}{12n} - \frac{1}{360n^3} + \dots\right).$$

It is known that if no terms of the series are taken, $n!$ is underestimated, if only the $\frac{1}{12n}$ term is taken, then $n!$ is overestimated, and so on. We wish to overestimate $n!/(\lambda n)!(\mu n)!$. This will be done if the numerator is overestimated and the denominator underestimated. Thus we may write

$$\frac{n!}{(\lambda n)!(\mu n)!} \leq \frac{(2\pi)^{1/2} n^{n+1/2} e^{-n} \exp\left(\frac{1}{12n}\right)}{(2\pi)^{1/2} \lambda n^{\lambda n+1/2} e^{-\lambda n} \exp\left(\frac{1}{12\lambda n} - \frac{1}{360(\lambda n)^3}\right) (2\pi)^{1/2} \mu n^{\mu n+1/2} e^{-\mu n} \exp\left(\frac{1}{12\mu n} - \frac{1}{360(\mu n)^3}\right)}$$

or

$$\frac{n!}{(\lambda n)!(\mu n)!} \leq \frac{1}{\sqrt{2\pi n \lambda \mu}} \cdot \frac{1}{\lambda^{\lambda n} \mu^{\mu n}} \exp\left(\frac{1}{12n} - \frac{1}{12\lambda n} - \frac{1}{12\mu n} + \frac{1}{360(\lambda n)^3} + \frac{1}{360(\mu n)^3}\right).$$

We wish to show that the exp term is less than or equal to one, or, which is the same thing, that its argument is less than or equal to zero.

One or the other of λ, μ is the greater. From symmetry, we may assume without loss of generality that it is λ , that is, $\lambda \geq \mu$. Then $\frac{1}{360(\lambda n)^3} \leq \frac{1}{360(\mu n)^3}$ and since μn is a positive integer, $\frac{1}{360(\mu n)^3} \leq \frac{1}{360\mu n}$. Further, $\frac{1}{12n} - \frac{1}{12\lambda n} \leq 0$, since $\lambda n \leq n$. Using these, we have

$$\left(\frac{1}{12n} - \frac{1}{12\lambda n} - \frac{1}{12\mu n} + \frac{1}{360(\lambda n)^3} + \frac{1}{360(\mu n)^3}\right) \leq \left(\frac{1}{12n} - \frac{1}{12\lambda n}\right) - \left(\frac{1}{12\mu n} - \frac{1}{180\mu n}\right) \leq 0.$$

This proves the upper bound (2). The lower bound is found similarly by underestimating the numerator and overestimating the denominator. No terms of the series are used for $n!$ and the $\frac{1}{12\lambda n}$ and $\frac{1}{12\mu n}$ for the denomi-

nator term. This gives directly

$$\frac{n!}{(\lambda n)! (\mu n)!} \geq \frac{1}{\sqrt{2\pi n \lambda \mu}} \cdot \frac{1}{\lambda^{\lambda n} \mu^{\mu n}} \exp -\left(\frac{1}{12\lambda n} + \frac{1}{12\mu n}\right).$$

The other lower bound with $\sqrt{\pi}/2$ in place of the exp term is obtained by noting first that unless both λn and μn are less than or equal to two, the argument of the exponential $\left(\frac{1}{12\lambda n} + \frac{1}{12\mu n}\right)$ is less than $\left(\frac{1}{12} + \frac{1}{36}\right) = \frac{1}{9}$. Now $\exp -\frac{1}{9} \geq \sqrt{\pi}/2$, and it is also readily verified that for the four cases where both λn and μn do not exceed two, namely (2,2), (2,1), (1,2) and (1,1), that the result is true. The worst case is (1,1) which just gives $\sqrt{\pi}/2$ for equality. Hence the result is true in general.

The upper and lower bounds (4) for the multinomial are found in exactly the same way as for the binomial.

The tail inequality for the binomial is formed by overestimating the tail using an infinite geometric series. This process is familiar (see, for example, Feller) with g replaced by the binomial coefficient. The inequality (6) is a special case of Chernoff's inequality which will be discussed later more generally.

A Lower Bound on the Tail of a Distribution

Let $\mu(s)$ be the logarithm of the moment-generating function of a distribution $F(x)$, and assume $\mu(s)$ exists in an interval with $s = 0$ in its interior. Then

$$dF(x) = e^{\mu(s)} e^{-sx} dG(x) \quad (1)$$

where $G(x)$ is the distribution of the tilted random variable obtained from $F(x)$ by the e^{sx} multiplying operation and normalization. $G(x)$ has its mean at $\mu'(s)$ and its variance is $\sigma^2 = \mu''(s)$.

By the Chebycheff inequality

$$G(\mu'(s) + \alpha/\sqrt{\mu''(s)}) - G(\mu'(s) - \alpha/\sqrt{\mu''(s)}) \geq 1 - \frac{1}{\alpha^2}$$

for any positive α . Now integrate equation (1) from $\mu'(s) - \alpha/\sqrt{\mu''(s)}$ to $\mu'(s) + \alpha/\sqrt{\mu''(s)}$. This gives

$$F(\mu'(s) + \alpha/\sqrt{\mu''(s)}) - F(\mu'(s) - \alpha/\sqrt{\mu''(s)}) = e^{\mu} \int_{\mu'(s) - \alpha/\sqrt{\mu''(s)}}^{\mu'(s) + \alpha/\sqrt{\mu''(s)}} e^{-sx} dG(x) \quad (2)$$

$\geq e^{\mu - s\mu' - \frac{1}{2}\alpha^2/\mu''(s)} (1 - \frac{1}{\alpha^2})$

This then, is a lower bound on the probability for the F distribution in a small interval in terms of the logarithm of the moment generating function.

If F is the convolution of n identical and independent distributions, each with $\mu(s)$ for its log moment generating function, then that for F itself is equal to $n\mu(s)$. The interval in question is then $2\alpha/\sqrt{n\mu''(s)}$ while the center position (for a fixed s) grows as $n\mu'(s)$.

If we integrate (1) from $-\infty$ to $\mu' + \alpha/\sqrt{\mu''}$ and assume $s \leq 0$ we obtain an underbound on the tail of the distribution F in the negative direction. This gives

$$F(\mu' + \alpha/\sqrt{\mu''}) \geq e^{\mu} \int_{-\infty}^{\mu' + \alpha/\sqrt{\mu''}} e^{-sx} dG(x)$$

$$\geq e^{\mu - s\mu' + s\alpha/\sqrt{\mu''}} \int_{-\infty}^{\mu' + \alpha/\sqrt{\mu''}} dG(x) \quad \begin{matrix} s \leq 0 \\ \alpha \geq 0 \end{matrix}$$

$$F(\mu' + \alpha/\sqrt{\mu''}) \geq (1 - \frac{1}{\alpha^2}) e^{\mu - s\mu' + s\alpha/\sqrt{\mu''}}$$

If F is the convolution of n identical distributions each with $\mu(s)$ as the logarithm of its moment generating function,

$$F(n(\mu' - \sigma\sqrt{\frac{\mu''}{n}})) \geq (1 - \frac{1}{\sigma}) e^{-n(\mu' - \sigma\mu + \sigma\sqrt{\frac{\mu''}{n}})}$$

Thus the argument of F approaches asymptotically for large n the argument $n\mu'$ appearing in the Chernoff upper bound. Likewise the exponent on the right (and the coefficient $1 - \frac{1}{\sigma}$ can also be included as a term in the exponent) approaches asymptotically the exponent in the Chernoff upper bound.

These inequalities may also be extended to the case where F is a convolution of not necessarily identical distributions with functions $\mu_i(s)$ ($i = 1, 2, \dots, n$). Then for F itself we have $\mu = \sum \mu_i$, $\mu' = \sum \mu'_i$ and $\mu'' = \sum \mu''_i$, and these may be substituted in (2) and (3). It is also evident that these same inequalities for $s \geq 0$ give a lower bound on the tail in the positive direction, that is, $1 - F(\mu' - \sigma\sqrt{\mu''}) = 0$.

Lower Bounds on Multinomial Tails and Terms*

Suppose we have a discrete distribution: a random variable can assume values $v_1 < v_2 < \dots < v_t$ with probabilities p_1, p_2, \dots, p_t . We wish to establish a lower bound on the size of term that can be found in a small interval when this distribution is convolved with itself n times (that is, jumps in the distribution of the sum of n independent variables, each with the given distribution). We first show the existence of a term having a certain size near the mean of the convolved distribution. To do this, the following lemma is first proved.

* Parts of these results were obtained in collaboration with Peter Elias.

Lemma: For any given n , we can find integers n_1, n_2, \dots, n_t such that

$$|n_i - p_i n| \leq 1 \quad (1)$$

$$\sum n_i = n \quad (2)$$

$$n \sum p_i v_i \leq \sum n_i v_i \leq n \sum p_i v_i + \Delta \quad (3)$$

where $\Delta = \min_i v_i + 1 - v_i$.

Proof: We first find a set of integers m_i which satisfy all the conditions except $\sum m_i v_i < n \sum p_i v_i + \Delta$, and will then derive from these the n_i . Choose m_t to be the first integer greater than $p_t n$. Set $m_t - p_t n = \delta_t$. Next, choose m_1 as the greatest integer less than $p_1 n$ and set $m_1 - p_1 n = \delta_1$. If $\delta_t - \delta_1 > 0$, take another m from the low end (i. e., m_2), the largest integer less than $p_2 n$, and then calculate $\delta_t + \delta_1 + \delta_2$ where $\delta_2 = m_2 - p_2 n$. If this is positive, proceed with p_3 , etc., until the accumulated sum of δ 's first becomes non-positive. When this occurs, terms are taken from the top end of the v range (p_{t-1}, p_{t-2} , etc.) until the accumulated sum of δ 's goes positive.

This process is continued, alternating from one end to the other as the sum of the δ 's changes sign, and eventually will end with some index k , having the property that all n_i for $i \leq k$ satisfy $n_i - p_i n = \delta_i \leq 0$ while for all n_i with $i \geq k$, we have $n_i - p_i n = \delta_i > 0$. At each stage of the operation, the total accumulated discrepancy satisfies $|\sum \delta_i| \leq 1$. This is true at the beginning, and arguing inductively at each stage we add a δ of absolute value less than or equal to one to an accumulated $\sum \delta_i$ of absolute value less than or equal to one and of opposite sign. This leads to the next accumulated sum also being less than or equal to one in absolute

value. Hence, when the last assignment of m_k is to be made, $|\sum \delta_i| \leq 1$.

If we let $m_k = n - \sum_{i \neq k} m_i$, then we satisfy $\sum m_i = n$ and also have

$$m_k = n - \sum_{i \neq k} (np_i + \delta_i)$$

$$= n - (n - np_k) - \sum_{i \neq k} \delta_i$$

$$= np_k + \theta$$

$$|\theta| \leq 1$$

Thus, $|\delta_k| \leq 1$ also.

Now since $\sum_{i=1}^t \delta_i = 0$, we have

$$-\sum_{i=1}^h \delta_i = \sum_{i=h+1}^t \delta_i$$

where h is the index of the largest negative δ_i , (either δ_{k-1} or δ_k).

Multiplying each side by v_h and using the monotone ordering of the v_i we obtain

$$-\sum_{i=1}^h \delta_i v_i \leq -\sum_{i=1}^h \delta_i v_h = \sum_{i=h+1}^t \delta_i v_h \leq \sum_{i=h+1}^t \delta_i v_i$$

Hence, using the end expressions in the above inequalities:

$$\sum_{i=1}^t \delta_i v_i \geq 0$$

and therefore

$$\sum_{i=1}^t m_i v_i = \sum_{i=1}^t (np_i + \delta_i) v_i$$

$$= \sum_{i=1}^t np_i v_i + \sum_{i=1}^t \delta_i v_i$$

$$= n \sum_{i=1}^t p_i v_i + \sum_{i=1}^t \delta_i v_i$$

$$= n \sum_{i=1}^t p_i v_i$$

Now starting with the m_i we can construct a set of n_i which satisfy all the conditions of the lemma. Note first that all the δ_i for $i > h$ are positive and for $i \leq h$ are negative. If we replace one of the lower m_i , say m_a ($a \leq h$), by the next larger integer $m_a + 1$ and simultaneously an m_b ($b > h$) by the next lower integer $m_b - 1$, we retain the properties that the errors in approximation satisfy $|\delta_i| \leq 1$ and that their sum be zero (or equivalently, $\sum m_i = n$). However, this reduces the value of $\sum m_i v_i$ by an amount $v_b - v_a$. Starting with the set of m_i just derived, we shall show how by interchanges of this type it is possible to go down from the value $\sum m_i v_i$ by steps none of which is larger than Δ , and eventually arrive at a sum less than or equal to $n \sum p_i v_i$. It will follow that in this sequence of operations there is a stage at which the third condition of the lemma obtains.

The series of steps is constructed as follows. Perform the interchange operation on m_h (the last negative m_i) and m_{h+1} . Since $v_{h+1} - v_h \leq \Delta$, the change in the sum due to this change is less than or equal to Δ . Now in place of this interchange consider that of h against $h+2$, or that of $h-1$ against $h+1$. The additional change in these cases over that just considered is clearly less than or equal to Δ , being indeed $v_{h+2} - v_{h+1}$ or $v_h - v_{h-1}$. The next stage would involve adding to one end or the other of the interval already taken. This again changes the sum from that previously obtained by not more than Δ . This process is continued until the ends of the range are reached, that is, v_t and v_1 are used in the interchange. These are now left in the changed state and the process is started again with m_h and m_{h+1} . Working outward from these eventually the numbers m_2 and $m_t - 1$ are used. These are then left

in the changed state (that is at $m_2 + 1$ and $m_t - 1$) and again the process started at m_h and $m_h + 1$. This procedure is continued until the permanently changed m 's from one end or the other reach m_h or $m_h + 1$ so that further steps of this type are not possible. The set of changed m_i 's, say m_i' , then existing have essentially the reverse property of the original m_i set: the corresponding δ_i' (that is $m_i' - p_i n$) satisfy $\delta_i' \geq 0$ for $i \leq h'$ for a certain h' . Hence, using essentially the same argument we used in proving (4), we can show that

$$\sum_{i=1}^t \delta_i' v_i \leq 0.$$

Thus this series of steps has at some stage given a set of integers n_i such that $0 \leq \sum_{i=1}^t n_i \delta_i' \leq \Delta$, namely, the integers at the stage just before this sum goes negative. For these n_i we have, equivalently,

$$n \sum_{i=1}^t p_i v_i \leq \sum_{i=1}^t n_i v_i \leq n \sum_{i=1}^t p_i v_i + \Delta.$$

This completes the proof of the lemma.

Returning now to the original problem, consider the term in the n th convolved distribution where the value v_i is taken n_i times ($i = 1, 2, \dots, t$), the n_i being those of the lemma. In the multinomial distribution this gives rise to a term of total probability

$$\binom{n}{n_i} \prod_i p_i^{n_i} \frac{1}{\sqrt{2\pi n} \prod_i \frac{n_i}{n}} \exp \left(-\sum \frac{1}{12} \frac{n_i}{n} + \sum n_i \log p_i - \sum n_i \log \frac{n_i}{n} \right) \quad (5)$$

This inequality is an application of the general inequality proved previously for multinomials. We now wish to simplify this making use of the fact that the n_i are close to $p_i n$; $|\delta_i| = |n_i - p_i n| \leq 1$. Consider the last terms in

the exponential:

$$\begin{aligned}
 \sum n_i \log p_i &= \sum n_i \log \frac{n_i}{n} = - \sum n_i \log \left(1 + \frac{\delta_i}{p_i n}\right) \\
 &= - \sum n_i \frac{\delta_i}{p_i n} \quad (\text{since } \log(1+x) \leq x) \\
 &= - \sum (p_i n + \delta_i) \frac{\delta_i}{p_i n} \\
 &= - \sum \frac{\delta_i^2}{p_i n} \quad (\text{since } \sum \delta_i = 0) \\
 &= - \frac{1}{n} \sum \frac{1}{p_i}
 \end{aligned}$$

The first exponential term can be estimated as follows.

$$- \sum \frac{1}{12 n_i} = - \frac{1}{12 n} \sum \frac{1}{p_i} \frac{p_i n}{n_i}$$

We now assume that, for each i , $p_i n \geq 1$ (in other words, that $n \geq p_{\min}^{-1}$).

It then follows that each $n_i \geq 1$ (since $|p_i n - n_i| \leq 1$ and n_i is an integer) and hence

$$\begin{aligned}
 p_i \frac{n}{n_i} &= \frac{n_i - \delta_i}{n_i} \\
 &\leq \frac{n_i + 1}{n_i} \\
 &= 1 + \frac{1}{n_i} \\
 &\leq 2
 \end{aligned}$$

Thus

$$\begin{aligned}
 - \sum \frac{1}{12 n_i} &\geq - \frac{1}{12 n} \sum 2 \frac{1}{p_i} \\
 &= - \frac{1}{6 n} \sum \frac{1}{p_i}
 \end{aligned}$$

Finally the coefficient in (5) can be underbounded as follows.

$$\left(\frac{n_1}{n}\right) = \left(\frac{1}{p_1}\right) \left(1 + \frac{\delta_1}{np_1}\right)^{-1/2}$$

$$\geq \left(\frac{1}{p_1}\right)^{-1/2} \exp\left(-\frac{1}{2} \sum \frac{\delta_1}{np_1}\right)$$

$$\geq \left(\frac{1}{p_1}\right)^{-1/2} \exp\left(-\frac{1}{2n} \sum \frac{1}{p_1}\right)$$

Collecting these various terms we have the following result:

Theorem: The sum of n independent random variables, each with the same discrete distribution, probability p_i of value v_i ($i = 1, 2, \dots, t$) ($v_1 \leq v_{i+1}$) has a term in the closed interval from $\sum p_i v_i$ to $\sum p_i v_i + \Delta$ where $\Delta = \max(v_{i+1} - v_i)$ and the term has a value at least $\frac{1}{\sqrt{2\pi n} \prod p_i} e^{-\frac{1}{2} \sum \frac{1}{p_i}}$, provided $n > p_{\min}^{-1}$.

This result may be generalized to give a term of such a distribution anywhere in the possible range. This is done by writing the distribution in terms of the tilted distribution; the sum of independent random variables with probabilities $q_i(s) = p_i e^{v_i s} / \sum p_i e^{v_i s}$. As we have seen previously, the distribution function of the original sum, $F_n(x)$, is related to that of the tilted distribution function, $G_n(x)$, by the equation

$$dF_n(x) = e^{\mu(s)} e^{-sx} dG_n(x)$$

The G_n distribution has a term in the interval $A = \mu'(s)$ to $A + \Delta$ since $\mu'(s)$ is the mean and the previous result applies. This gives a term in the F_n distribution, to the amount stated in the following.

Theorem: The sum of n independent random variables, each with the same discrete distribution, probability p_i of value v_i , ($v_1 < v_{i+1}$) ($i = 1, 2, \dots, t$), has a term in the closed interval from A to $A + \Delta$ where $\Delta = \max_i (v_{i+1} - v_i)$ and $n v_{\min} \leq A \leq n v_{\max}$. The term will have a

value at least

$$\frac{1}{2\pi n \prod q_i(s)} e^{-A|s|} - \frac{5}{3n} \sum \frac{1}{q_i(s)} e^{\mu(s) - s\mu'(s)}$$

where $q_i(s) = p_i e^{v_i s} / \sum p_i e^{v_i s}$ and s is chosen to make $A = \sum q_i(s) v_i$,
 and provided $n > q_1(s)$. The last term is the Chernoff bound with
 $\mu(s) = \log \sum p_i e^{v_i s}$, $\mu'(s) = A$.

A Combinatorial Theorem

Theorem: Suppose we have a set of objects S_1, S_2, \dots, S_n and a number of numerically valued properties (functions) for the objects P_1, P_2, \dots, P_d . These are non-negative $P_i(S_j) \geq 0$ and we know the averages of these properties over the objects:

$$\frac{1}{n} \sum_j P_i(S_j) = A_i \quad i = 1, 2, \dots, d$$

Then there exists an object S_p for which

$$P_i(S_p) \leq d A_i \quad i = 1, 2, \dots, d$$

More generally given any set of $K_i > 0$ satisfying

$$\sum_{i=1}^d \frac{1}{K_i} \leq 1$$

then there exists an object S_p

$$P_i(S_p) \leq K_i A_i \quad i = 1, 2, \dots, d$$

Proof: The second part implies the first by taking $K_i = d$. To prove the second part let N_i be the number of objects for which $P_i(S) > K_i A_i$. Now $A_i > \frac{1}{n} N_i K_i A_i$ (since all S 's have P_i values ≥ 0).

$$\text{Hence } N_i < \frac{n}{K_i}$$

The total number of objects M violating any of the conditions is less than or equal to the sum of the individual N_i

$$M < n \sum \frac{1}{K_i} \leq n \quad \text{using} \quad \sum \frac{1}{K_i} \leq 1$$

Hence there is at least one object not violating any of the conditions.

Some Results on DeterminantsThe root of a determinant equation.

Lemma: Given $f_{ij}(\omega)$ ($i, j = 1, 2, \dots, d$) continuous functions of ω in the range $a \leq \omega \leq b$ and in this range $f_{ij}(\omega) \geq 0$,

$\sum_j f_{ij}(\omega) > 0$, $f_{ij}(a) < \frac{1}{d}$, $f_{ij}(b) > \frac{1}{d}$, then there exists W , $a \leq W \leq b$

and a set of $X_i \geq 0$, $\sum X_i = 1$, such that

$$|f_{ij}(W) - \delta_{ij}| = 0$$

$$\sum_i X_i f_{ij}(W) = X_j$$

Proof: Consider the d dimensional region R whose points are (X_1, \dots, X_d, W) , where $X_i \geq 0$, $\sum X_i = 1$, $a \leq W \leq b$. This is a topological image of a sphere and its interior. For a fixed W in the range from a to b , consider the continuous mapping

$$X_j \rightarrow Y_j = \frac{\sum_i X_i f_{ij}(W)}{\sum_{ij} X_i f_{ij}(W)} \quad W \rightarrow V = \begin{cases} V_1 = W + 1 - \sum_{ij} f_{ij}(W) X_i \\ \text{if } a \leq V_1 \leq b \\ a \text{ if } V_1 < a \\ b \text{ if } V_1 > b \end{cases}$$

Note that the denominator for Y_j does not vanish because of our assumption that $\sum_j f_{ij}(\omega) > 0$ and hence the Y_j are well defined. Also the Y_j are

non-negative and $\sum_j Y_j = 1$. Finally $a \leq V \leq b$. Hence this maps points

(X_1, W) in R continuously into points (Y_1, V) in R . Consequently, by the Brouwer fixed point theorem there exists a point (X_1, W) which is mapped into itself, that is, a point for which $\sum_i X_i f_{ij}(W) = X_j \sum_{ij} X_i f_{ij}(W)$.

$W = V$. The value of W for the fixpoint clearly is not a or b since these points are moved upward or downward by our assumptions. Hence for the fixpoint we have $W = W + 1 - \sum_{ij} f_{ij}(W) X_i$ or $\sum_{ij} f_{ij}(W) X_i = 1$. It follows

that for the fixpoint

$$\sum_{ij} f_{ij}(W) X_i = X_j$$

$$|f_{ij}(W) - \delta_{ij}| = 0$$

Let the elements a_{ij} of a matrix be non-negative. Suppose there is an eigen vector A_i all of whose components are positive, $A_i > 0$, and the corresponding characteristic value is λ_0 . We will show that for any other characteristic value λ_1 we have $|\lambda_1| \leq \lambda_0$. Let B_i be a characteristic vector for λ_1 where we adjust the length of this vector as follows. Choose its length in such a way that $A_i - |B_i| \geq 0$ for all i and the equality holds for at least one i , say $i = h$, so that $A_h = |B_h|$. It is clear that this can be done since with zero length all components of B are less than those of A and increasing continuously, eventually a first one of the $|B_i|$ reaches its corresponding A_i . We now have

$$\sum_i A_i a_{ij} = \lambda_0 A_j \quad (1)$$

$$\sum_i B_i a_{ij} = \lambda_1 B_j \quad (2)$$

$$\sum_i |B_i| a_{ij} \geq |\lambda_1| |B_j| \quad (3)$$

Subtracting these equations for $j = h$

$$\begin{aligned} \sum_i (A_i - |B_i|) a_{ih} &\leq \lambda_0 A_h - |\lambda_1| |B_h| \\ &= (\lambda_0 - |\lambda_1|) A_h \end{aligned} \quad (4)$$

All terms in the sum at the left are non-negative and also A_h is definitely positive. It follows that $\lambda_0 - |\lambda_1| \geq 0$.

The derivative of the eigenvalue of a matrix.

Suppose we have the square matrix $(a_{ij}(s))$ where the elements are differentiable functions of a parameter s . Let $\nu = \nu(s)$ be an eigenvalue with corresponding eigen vector $A_i = A_i(s)$ and eigen vector $B_j = B_j(s)$ for the transposed matrix. Thus

$$|a_{ij}(s) - \nu(s)\delta_{ij}| = 0 \quad (1)$$

$$\sum_i A_i a_{ij} = \nu A_j \quad (2)$$

$$\sum_j B_j a_{ij} = \nu B_i \quad (3)$$

Theorem:

$$\gamma'(s) = \frac{\sum_{ij} A_i a'_{ij} B_j}{\sum_i A_i B_i}$$

To prove this, differentiate (2) with respect to s :

$$\sum_i A'_i a_{ij} + \sum_i A_i a'_{ij} = \gamma'_j A_j + \gamma A'_j .$$

Now multiply by B_j and sum on j

$$\sum_{ij} A'_i a_{ij} B_j + \sum_{ij} A_i a'_{ij} B_j = \gamma' \sum_j A_j B_j + \gamma \sum_j A'_j B_j .$$

Using (3) in the first term cancels the last term on the right, giving the desired result

$$\sum_{ij} A_i a'_{ij} B_j = \gamma' \sum_j A_j B_j .$$

Upper and Lower Bounds for Powers of a Matrix with Non-negative Elements

We frequently have to deal with the n^{th} power of a matrix whose elements are $\beta_{ij} \geq 0$. We denote the ij element of this n^{th} power by $\beta_{ij}^{(n)}$. We are concerned here with the case where the corresponding graph has the property that it is possible to go from any node i to any other j by a finite sequence $\beta_{ia}, \beta_{ab}, \dots, \beta_{gj}$ where all the β 's in this series are positive. This means that the graph consists of one ergodic or periodic set in the usual Markoff analysis. The non-negative condition on the β_{ij} insures the existence of a real eigenvalue v_0 which is a solution of the determinant equation $\left| \beta_{ij} - v\delta_{ij} \right| = 0$. Further, this v_0 dominates in absolute value any other eigenvalue v_1 , that is, $v_0 \geq |v_1|$.

Corresponding to root v_0 there will exist right and left eigen-vectors for the matrix

$$\begin{aligned} \sum_i \beta_{ij} &= v_0 A_j \\ \sum_j \beta_{ij} B_j &= v_0 B_i \end{aligned} \quad (1)$$

The conditions $\beta_{ij} \geq 0$ imply that all the A_i be the same sign (or vanish) and all the B_i be the same sign (or vanish). In both cases we take them to be positive (multiply by -1 if necessary). In the case satisfying the graphical condition it is easily seen that all A_i and all B_i are then actually positive (none vanish).

Theorem: Under the conditions above, i. e. $\beta_{ij} \geq 0$ and any state accessible from any other through a finite sequence of non-vanishing transitions, the element $\beta_{ij}^{(n)}$ of $\left\| \beta_{ij} \right\|^n$ is bounded by

$$\beta_{ij}^{(n)} \leq \frac{B_i}{B_j} v_0^n \leq v_0 \beta_{ij}^{(n-d)} v_0^n$$

where β_{\min} is the smallest (non-vanishing) β_{ij} , and d is an integer such that there is a path from any state i to any state j with not more than d steps ($d-1$ intermediate states). Furthermore, there will exist $k > 0$ and n_0

such that

$$\beta_{ij}^{(n)} \geq k \frac{B_j}{B_i} v_0^n \quad n \geq n_0$$

provided either (1) for some n_1 , $\beta_{ij}^{(n_1)} > 0$ for all i, j or (2) the state diagram has no recurrent subsets (the greatest common divisor of closed path lengths is 1).

Proof: The first inequality is proved easily by induction on n .

For $n = 0$,

$$\beta_{ij}^{(0)} = \delta_{ij} \leq B_i/B_j$$

since for $i \neq j$, the right member is positive and $\delta_{ij} = 0$, while for $i = j$,

$\delta_{ii} = 1$ and the right member is one.

Now supposing the inequality to hold for n we prove it for $n+1$.

$$\beta_{ij}^{(n+1)} = \sum_s \beta_{is} \beta_{sj}^{(n)}$$

$$\leq \sum_s \beta_{is} B_j^{-1} B_s v_0^n$$

$$= B_j^{-1} v_0^n \sum_s \beta_{is} B_s$$

$$= B_j^{-1} v_0^n v_c B_i$$

This is the corresponding inequality for $n+1$, concluding the proof.

The second inequality, that $B_i/B_j \leq (v_0/\beta_{\min})^d$ is shown as follows.

From (1), let some β_s be positive then

The Number of Sequences of a Given Length

Suppose a number of letters are available whose lengths (or durations) are a_1, a_2, \dots, a_g and we wish a bound on the number $N(\ell)$ sequences of total length ℓ . Here it is assumed that any sequence of letters is allowed. $N(\ell)$ satisfies the difference equation

$$N(\ell) = N(\ell - a_1) + N(\ell - a_2) + \dots + N(\ell - a_g) \quad \ell > 0$$

as we see by noting that each sequence of length ℓ must end in one or another of the available letters. Furthermore, the boundary conditions may be taken to be $N(\ell) = 0$ for $\ell < 0$ and $N(0) = 1$. Associated with the difference equation is the following characteristic equation:

$$1 - X^{-a_1} + X^{-a_2} + \dots + X^{-a_g} = 0.$$

Since all the a_i are positive and real, the right-hand member is a strictly monotone decreasing function of X and varies from ∞ to 0 when X goes from 0 to ∞ . Consequently, the characteristic equation has a unique positive real root W .

Theorem: $N(\ell) \leq W^\ell$.

To prove this, note first that W^ℓ satisfies the difference equation since this results on multiplying the characteristic equation (with X replaced by W) by W^ℓ . With regard to the boundary conditions, $W^0 = 1 = N(0)$ and $W^\ell > 0 = N(\ell)$ when $\ell < 0$. Let a be the smallest of a_1, a_2, \dots, a_g . Then it is possible to proceed by a kind of induction of steps of ℓ (each of length a) to show that the dominance of W^ℓ over $N(\ell)$ continues for all ℓ . In fact, suppose that for $\ell \leq \ell_1$ we have $N(\ell) \leq W^\ell$. Then for ℓ in the range $\ell_1 \leq \ell \leq \ell_1 + a$

$$\begin{aligned} N(\ell) &= N(\ell - a_1) + N(\ell - a_2) + \dots + N(\ell - a_g) \\ &\leq W^{\ell - a_1} + W^{\ell - a_2} + \dots + W^{\ell - a_g} \\ &\leq W^\ell. \end{aligned}$$

Since the inequality is true for $\ell \leq 0$, it follows that it is true for all ℓ .

A more general problem of the same sort relates to sequences which are subject to a finite state set of constraints. Thus, suppose there are d states and that in state i , letters of lengths ℓ_{aij} are permitted,

leading to state j . The index a ranges over the different letters going from state i to state j and j ranges over the different states which can follow state i . Now let $N_{ij}(\ell)$ be the number of sequences which are possible and which start in state i , end in state j and are of length ℓ . These quantities are readily seen to satisfy the difference equations

$$N_{ij}(\ell) = \sum_{a,k} N_{ik}(\ell - \ell_{akj}) \quad \ell > 0 \quad (1)$$

$$N_{ij}(\ell) = 0 \quad \ell < 0$$

The corresponding characteristic equations are

$$A_j = \sum_{a,i} A_i W^{-\ell_{a ij}} \quad (2)$$

Let W be the largest real root (there is a positive real root by a previous result based on the fix point theorem) of the determinant equation:

$$\left| \sum_a W^{-\ell_{a ij}} - \delta_{ij} \right| = 0$$

and let A_i be a corresponding (positive) solution of (2). We will assume the graph of the constraints is fully connected so it is possible to go from any state to any other. Then all the A_i are positive (none vanish).

We will now show that the number of sequences of length ℓ starting in state i and ending in j , $N_{ij}(\ell)$, is bounded by

$$N_{ij}(\ell) \leq \frac{A_j}{A_i} W^\ell$$

This is certainly true for $\ell < 0$ and also for $\ell = 0$ since then both sides are one if $i = j$, and otherwise the left side is zero with the right positive. We now proceed by the inductive type process as before, assuming the inequality out to some ℓ_1 and then show it follows for ℓ out to ℓ_1 plus the minimum $\ell_{a ij}$.

$$\begin{aligned} N_{ij}(\ell) &= \sum_{a,s} N_{is}(\ell - \ell_{asj}) \\ &\leq \sum_{a,s} \frac{A_s}{A_i} W^{\ell - \ell_{asj}} \quad \ell \leq \ell_1 + \min \ell_{a ij} \\ &= \frac{W^\ell}{A_i} \sum_{a,s} A_s W^{-\ell_{asj}} \end{aligned}$$

(continued next page)

$$= W^{\ell - \frac{A_1}{A_1}}$$

Thus the inductive step carries the inequality up to $\ell = \ell_1 + \min \ell_{aij}$, and hence it is true for all ℓ .

An Alternative Proof that $N(\ell) \leq W^\ell$

Consider the case of a sequence of letters of different lengths a_1, a_2, \dots, a_g with no constraints. We wish to prove that $N(\ell) \leq W^\ell$, where W satisfies $\sum_i W^{-a_i} = 1$. Assume, in contradiction, that for some ℓ , $N(\ell) > W^\ell$. Then, since $N(0) \leq W^0$, there is a greatest lower bound of ℓ 's, say ℓ^* , for which the theorem fails. In the interval $\ell^* \leq \ell \leq \ell^* + \frac{1}{2} a_{\min}$ there must be an ℓ , say ℓ_1 , for which the theorem fails (a_{\min} is the smallest a_i). Subdivide the sequences of length ℓ_1 into subsets according to the first letter. Let the fractional number in the subset beginning with the letter i be f_i ($i = 1, 2, \dots, g$). Choose the subset for which $a_i^{-1} \log f_i^{-1}$ is a minimum. In a sense, this means the subset which conveys the least information, $\log f_i^{-1}$, per unit time in its first letter. The minimum value of $a_i^{-1} \log f_i^{-1}$ among the different subsets is less than or equal to $\log W$. To see this, suppose, in contradiction, that for all i , $a_i^{-1} \log f_i^{-1} > \log W$. Then $f_i < W^{-a_i}$ and, summing on i , $1 = \sum f_i < \sum W^{-a_i} = 1$, a contradiction. Hence the subset chosen will have $a_i^{-1} \log f_i^{-1} \leq \log W$, or $f_i \geq W^{-a_i}$. If we delete the first letter from all sequences in this subset, we are left with a set of more than $W^{\ell_1 - a_i}$ sequences of length $\ell_1 - a_i$. Thus $N(\ell_1 - a_i) > W^{\ell_1 - a_i}$. Since $\ell_1 - a_i < \ell^*$, this contradicts the assumption that ℓ^* was the greatest lower bound of ℓ 's for which the theorem fails. Hence the theorem is true for all ℓ .

Characteristic for a Language with Independent Letters

Suppose we have a stochastic process generating a language consisting of a sequence of independent letters. These letters are all chosen with the probabilities p_i for letter i , $i = 1, 2, \dots, g$. We consider sequences of n such letters, that is, words of length n in the language. Suppose that all such words are arranged in order of decreasing probability from the most probable one, consisting of a sequence of n most probable letters, down to the sequence of n least probable letters. The logarithm of the probability of any particular word is (because of the independence of letters) the sum of the logarithms of the probabilities of the individual letters. Thus, the logarithm of the probability of a word is a random variable which is the sum of n independent random variables each with the same distribution function. We may, therefore, apply previous results concerning the tails of such a distribution to estimate the probability in our monotone sequence of all words beyond a certain point.

The distribution of $\log p^{-1}$ for a single letter will have a moment generating function

$$\begin{aligned} \psi(s) &= \sum_i p_i e^{-s \log p_i} \\ &= \sum_i p_i^{1-s} \end{aligned}$$

Hence

$$\begin{aligned} \mu(s) &= \log \sum_i p_i^{1-s} \\ \mu'(s) &= \frac{\sum_i p_i^{1-s} \log p_i^{-1}}{\sum_i p_i^{1-s}} \end{aligned} \quad (1)$$

Our upper bound on the tail of a distribution then shows that the total probability P_T of all sequences whose individual probability P satisfies

$$\frac{1}{n} \log P \leq \mu'(s) = \frac{\sum_i p_i^{1-s} \log p_i^{-1}}{\sum_i p_i^{1-s}} \quad (2)$$

is bounded by

$$\frac{1}{n} \log P_T \leq \mu(s) - s\mu'(s) = \log \sum_i p_i^{1-s} + \frac{\sum_i p_i^{1-s} \log p_i}{\sum_i p_i^{1-s}} \quad (3)$$

This last expression as well as (1), can be written more compactly in terms of a new set of probabilities $q_i(s)$ defined as follows:

$$q_i(s) = \frac{p_i^{1-s}}{\sum_i p_i^{1-s}}$$

The relations (2) and (3) now become, after some manipulation,

$$\int_n \log \text{Prob} \left[\frac{1}{n} \log P \leq \sum_i q_i(s) \log p_i^{-1} \right] \leq \sum_i q_i(s) \log \frac{p_i}{q_i(s)} \quad (4)$$

This is one of the results we desire, an overbound on the tail of the distribution of probability for sequences.

We now desire a similar bound on the number of sequences whose probability is greater than P . To this end, consider constructing all sequences of length n giving each letter probability $\frac{1}{g}$ (instead of the probabilities p_i they actually have). We again consider the distribution of the sum of the logarithms of the probabilities (using the original p_i values) for the letters in a word. Note that the sequences arranged in monotone order are in the same order as previously. Under these new conditions the moment generating function $\nu_1(s)$ and its logarithm $\mu_1(s)$ are given by

$$\nu_1(s) = \sum_i \frac{1}{g} p_i^{-s}$$

$$\mu_1(s) = \log \sum_i p_i^{-s} - \log g$$

$$\mu_1'(s) = \frac{\sum_i p_i^{-s} \log p_i^{-1}}{\sum_i p_i^{-s}}$$

The total probability P_2 of all sequences in the tail of the distribution beyond the sequences whose individual probability P satisfies

$$\frac{1}{n} \log P \leq \mu_1(s) = \frac{\sum_i p_i^{-s} \log p_i^{-1}}{\sum_i p_i^{-s}} \quad (5)$$

will be bounded by

$$\frac{1}{n} \log P_2 \leq \mu_1(s) - s\mu_1(s) = \log \sum_i p_i^{-s} + \frac{\sum_i p_i^{-s} \log p_i^s}{\sum_i p_i^{-s}} = \log g.$$

We note first that in this modified probability system (each letter with probability $\frac{1}{g}$) all sequences have probability $\frac{1}{g^n}$ and consequently the number of sequences N_2 in the tail whose total probability is P_2 is precisely $P_2 g^n$. Hence the number N_2 in the tail is bounded by

$$\begin{aligned} \frac{1}{n} \log N_2 &= \frac{1}{n} \log P_2 g^n = \frac{1}{n} \log P_2 + \log g \\ &\leq \log \sum_i p_i^{-s} + \frac{\sum_i p_i^{-s} \log p_i^s}{\sum_i p_i^{-s}} \end{aligned}$$

In order to compare this result with the preceding one (4), we must identify the points at which the tails of the distributions are cut off. This can be done by equating the probabilities P of the individual sequences at the cutoff point. Thus, using (1) and (5) and writing s_1 in the latter in place of s we have

$$\frac{\sum_i p_i^{1-s} \log p_i^{-1}}{\sum_i p_i^{1-s}} = \frac{\sum_i p_i^{-s_1} \log p_i^{-1}}{\sum_i p_i^{-s_1}}$$

This is obviously satisfied by $1-s = -s_1$, and since $\mu''(s) > 0$ the left term is a strictly monotone function of s and therefore this solution is unique.

The number of sequences N_2 now becomes, in terms of the s involved in (1) and (4),

$$\frac{1}{n} \log N_2 \leq \log \sum_i p_i^{1-s} + \frac{\sum_i p_i^{1-s} \log p_i^{s-1}}{\sum_i p_i^{1-s}}.$$

Again using the $q_i(s)$ to simplify

$$\frac{1}{n} \log N_2 \leq \sum_i q_i(s) \log q_i(s)^{-1} \quad (6)$$

Both the bounds (4) and (6) are also the limiting values approached by $\frac{1}{n} \log P_T$ and $\frac{1}{n} \log N_2$ as $n \rightarrow \infty$. This follows from remarks concerning the tails of distributions made in an earlier section. Thus the reliability curve of a source of the type we are discussing here with independent letters may be written in parametric form as follows:

$$E(s) = -\sum_i q_i(s) \log \frac{p_i}{q_i(s)} = (1-s)\mu' \quad (7)$$

$$R(s) = \sum_i q_i(s) \log q_i(s)^{-1} = 1 - (1-s)\mu' \quad (8)$$

$$\text{where } q_i(s) = \frac{p_i^{1-s}}{\sum_i p_i^{1-s}} \quad (9)$$

The parameter s in these equations is related to the slope of the reliability curve. In fact, we note that

$$\frac{dE}{dR} = \frac{dE/ds}{dR/ds} = \frac{-\mu'(s) + s\mu''(s) + \mu'(s)}{\mu'(s) + (1-s)\mu''(s) - \mu'(s)} = \frac{s}{1-s}$$

Thus, as s increases from 0 to 1, the slope increases monotonically from 0 to ∞ . It is interesting that at $s = 1$ the formulas (7), (8) become

$$E(1) = \frac{1}{d} \sum \log p_i + \log d$$

$$R(1) = \log d$$

The Probability of Error in Optimal Codes

A problem of importance in information theory is that of studying the behavior of signaling codes that may be used in encoding an information source for a noisy channel and, in particular, the probability of error for the optimal code. This paper is concerned with estimating this probability of error under fairly general conditions.

We will find that, to a large extent, the problem can be divided into two parts. First, there is a problem relating to the information source only (not involving the channel) which involves estimating the probability of error when the source is encoded into a simple standard noiseless channel. The study of this question leads to a certain function which we call the reliability characteristic for the source and which determines, in a certain asymptotic sense when the code blocks are long, how rapidly the probability of error approaches zero. Second, there is a problem relating to the channel only. This leads to a function describing, in a sense, the coding behavior of the channel with regard to probability of error when the code blocks are long. Our final and most basic results show how the two functions may be combined to give optimal behavior (or bounds on optimal behavior) when the source is encoded into the channel.

We will first clarify our terminology, since various writers have used some of the terms involved with quite different meanings. For the most part, we will restrict ourselves to a finite, discrete, memoryless channel. Such a channel is specified by a transition probability matrix $\|p_i(j)\|$. Here $p_i(j)$ is the probability that if input symbol i is used, the output will be j and we have

$$\sum_j p_i(j) = 1.$$

Matrices satisfying the conditions that all elements are nonnegative and the row sums are unity occur often in probability and are called stochastic matrices.

The input symbols to the channel will be called the input letters, the set of these the input alphabet. The output symbols of the channel will be called the output letters and the set of these the output alphabet.

A channel is often conveniently represented by a line diagram of the type shown in Fig. 1.

The channel being memoryless means that successive operations are independent. If the input letters i and j are used, the probability of output letters k and ℓ will be $p_i(k)p_j(\ell)$. A sequence of input letters will be called an input word, a sequence of output letters an output word. A collection of M input words all of length n will be called a block code of length n . $R = 1/n \log M$ will be called the input rate for this code. Unless otherwise specified, a code will mean such a block code.

A detection system for a code is a method of interpreting output words as input words, that is, an association or mapping of one of the input words of the code for every output word of length n . The probability of error for a particular input word is the probability, if this input is used, that it will be interpreted incorrectly. It is, therefore, the probability of that input word being received as an output word which is not detected as the input word. The probability of error for a code is the average probability of error for all input words in the code. An optimal code of length n is one which minimizes this probability of error (when using its best detection system). These input words u_1, u_2, \dots, u_M need not all be different.

Our main problem is to estimate for a general channel upper and lower bounds on the probability of error for an optimal code as a function of the length of the code n and the rate of transmission R . The ideal solution would be to find a simple explicit formula for the probability of error in an arbitrary channel as a function of the rate of transmission R and the length of the code words n . This is probably too much to hope for in view of the diophantine complexities of optimal codes. Barring such a complete solution, one may still hope for upper and lower bounds on P_e and perhaps results relating to its asymptotic behavior when n is large. Most of the present paper is devoted to this type of result.

In studying the asymptotic behavior, it will appear that P_e , for a fixed rate R and a given channel, varies approximately exponentially with n . For this reason it is convenient to introduce a new term. If a device or a system has a probability P_e of making an error, we shall call $-\log P_e$

the reliability of the device or system. We have just said in effect that for large n the reliability for optimal codes varies essentially linearly with n , that is, as $E(R) \cdot n$, where R is the rate for the code. More precisely, we define $E(R)$ as follows:

$$E(R) = \limsup_{n \rightarrow \infty} -\frac{1}{n} \log P_e \text{ opt}$$

We will call $E(R)$ the reliability characteristic of the channel and attempt to evaluate it, or where we cannot do this, at least place upper and lower bounds on it.

The writer feels that the quantity we have defined as reliability will, in many cases, turn out to be the most appropriate way of measuring a probability of error. In previous work by von Neumann on unreliable neuron-type elements and by E. P. Moore and the writer on unreliable relays, the quantity $-\log P_e$ entered significantly and was the more natural way to describe some of the results. In both these cases the reliability varied rather simply with the redundancy of the error-correcting systems. It is a little like measuring gain on a db scale or ion concentration on a pH scale. While actually little more than a change in scale, the use of these units of reliability in the coding case throws the results into a much more natural and illuminating perspective.

If we have two given channels, it is possible to form a single channel from them in two natural ways which we call the sum and product of the two channels. The sum of two channels is the channel formed by using inputs from either of the two given channels with the same transition probabilities to the set of output letters consisting of the logical sum of the two output alphabets. Thus the sum channel is defined by a transition matrix formed by placing the matrix of one channel below and to the right of that for the other channel and filling the remaining two rectangles with zeros. If $\|p_i(j)\|$ and $\|p'_i(j)\|$ are the individual matrices, the sum has the following matrix:

$$\begin{array}{ccccccccc} p_1(1) & . & . & . & p_1(r) & 0 & . & . & . & 0 \\ \vdots & & & & \vdots & \vdots & & & & \vdots \\ p_t(1) & . & . & . & p_t(r) & 0 & . & . & . & 0 \\ 0 & . & . & . & 0 & p'_1(1) & . & . & . & p'_1(r') \\ \vdots & & & & \vdots & \vdots & & & & \vdots \\ 0 & . & . & . & 0 & p'_t(1) & . & . & . & p'_t(r') \end{array}$$

The product of two channels is the channel whose input alphabet consists of all ordered pairs (i, i') where i is a letter from the first channel alphabet and i' from the second, whose output alphabet is the similar set of ordered pairs of letters from the two individual output alphabets and whose transition probability from (i, i') to (j, j') is $p_i(j)p_{i'}(j')$.

Sum of two channels with independent inputs

The writer feels that the quantity we have defined as reliability, in many cases, turns out to be the most appropriate way of measuring a probability of error. In previous work by von Neumann on unreliable neuron-type elements and by J. L. Moore and the writer on unreliable relays, the quantity $10 \log_{10}$ entered signal intensity and was the more natural way to describe some of the results. In both cases the reliability varied rather simply with the redundancy of the error-correcting system. It is a little like measuring gain on a dB scale or loss concentration on a pH scale. This is actually little more than a change in scale, the use of these units of reliability in the coding case throws the results into a much more natural and illuminating perspective.

If we have two given channels, it is possible to form a single channel from them in two natural ways which we call the sum and product of the two channels. The sum of two channels is the channel formed by using inputs from either of the two given channels with the same transition probabilities to the set of output letters consisting of the logical sum of the two output alphabets. Thus the sum channel is defined by a transition matrix formed by placing the matrix of one channel below and to the right of that for the other channel and filling the remaining two rectangles with zeros. If $\|p_i(j)\|$ and $\|p_{i'}(j')\|$ are the individual matrices, the sum has the following matrix:

$$\begin{bmatrix} p_i(j) & 0 & \dots & 0 \\ 0 & p_{i'}(j') & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & p_{i'}(j') \end{bmatrix}$$

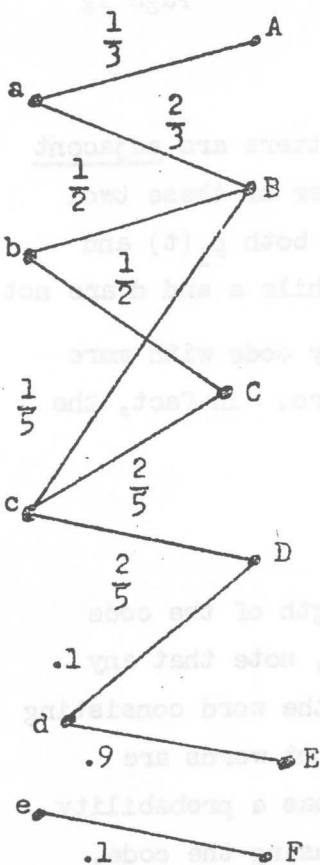


Fig. 1

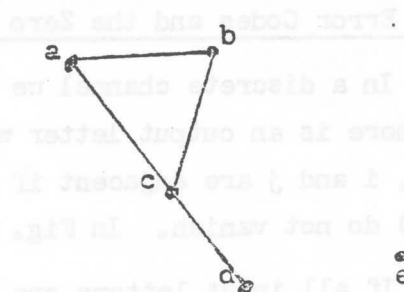


Fig. 3

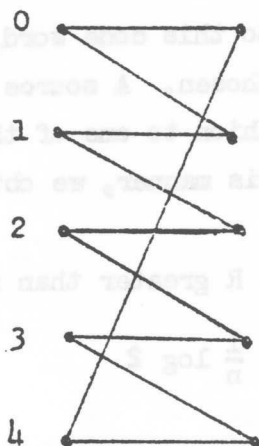


Fig. 2

Zero Error Codes and the Zero Error Capacity C_0

In a discrete channel we will say that two input letters are adjacent if there is an output letter which can be caused by either of these two. Thus, i and j are adjacent if there exists a t such that both $p_i(t)$ and $p_j(t)$ do not vanish. In Fig. 1, a and c are adjacent, while a and d are not.

If all input letters are adjacent to each other, any code with more than one word has a probability of error greater than zero. In fact, the probability of error satisfies

$$P_e \geq \frac{M-1}{M} p_{\min}^n$$

where p_{\min} is the smallest among the $p_i(j)$, n is the length of the code and M is the number of words in the code. To prove this, note that any two words have a possible output word in common, namely the word consisting of the sequence of common output letters when the two input words are compared letter by letter. Each of the two input words has a probability at least p_{\min}^n of producing this common output word. In using the code, the two particular input words will each occur $\frac{1}{M}$ of the time and will cause the common output $\frac{1}{M} p_{\min}^n$ of the time. This output can be decoded in only one way. Hence at least one of these situations leads to an error. This error, $\frac{1}{M} p_{\min}^n$, is assigned to this code word, and from the remaining $M-1$ code words another pair is chosen. A source of error to the amount $\frac{1}{M} p_{\min}^n$ is assigned in similar fashion to one of these, and this is a disjoint event. Continuing in this manner, we obtain a total of $\frac{M-1}{M} p_{\min}^n$ probability of error.

It follows that for any rate R greater than zero, (i.e. $M \geq 2$)

$$-\frac{1}{n} \log P_e \leq \log p_{\min}^{-1} + \frac{1}{n} \log 2$$

$$E \leq \log p_{\min}^{-1}$$

If it is not true that the input letters are all adjacent to each other, it is possible to transmit at a positive rate with zero probability of error. The least upper bound of all rates which can be achieved with zero probability of error will be called the zero error capacity of the channel and denoted by C_0 . If we let $M_0(n)$ be the largest number of words in a code of length n , no two of which are adjacent, then C_0 is the least upper bound of the numbers $\frac{1}{n} \log M_0(n)$ when n varies through all positive integers. An interesting problem which has not been completely

solved is that of evaluating C_0 for an arbitrary channel.

One might expect that C_0 would be equal to $\log M_0(1)$, that is, that if we choose the largest possible set of non adjacent letters and form all sequences of these of length n , then this would be the best error free code of length n . This is not, in general, true, although it holds in many cases, particularly when the number of input letters is small. The first failure occurs with five input letters with the channel in Fig. 2. In this channel, it is possible to choose at most two independent letters, for example 0 and 2. Using sequences of these, 00, 02, 20, and 22 we obtain four words in a code of length two. However, it is possible to construct a code of length two with five members no two of which are adjacent as follows: 00, 12, 24, 31, 43. It is readily verified that no two of these are adjacent. Thus, C_0 for this channel is at least $\frac{1}{2} \log 5$.

No method has been found for determining C_0 for the general discrete channel, and this we propose as an important unsolved problem in coding theory. We shall develop a number of results which enable one to determine C_0 in many special cases, for example, in all channels with five or less inputs with the single exception of the channel of Fig. 2 (or channels equivalent in adjacency structure to it). We will also develop some general inequalities enabling one to estimate C_0 quite closely in most cases.

It may be seen, in the first place, that the value of C_0 depends only on which input letters are adjacent to each other. Let us define an adjacency matrix for a channel, A_{ij} , as follows.

$$A_{ij} = \begin{cases} 1 & \text{if input letter } i \text{ is adjacent to } j \text{ or if } i = j \\ 0 & \text{otherwise} \end{cases}$$

Suppose two channels have the same adjacency matrix (possibly after renumbering the input letters of one of them.) Then it is obvious that a zero error code for one will be a zero error code for the other and, hence, that the zero error capacity C_0 for one will also apply to the other.

The adjacency structure contained in the adjacency matrix can also be represented as a linear graph. Construct a graph with as many vertices as there are input symbols, and connect two distinct vertices with a line or branch of the graph if the corresponding input letters are adjacent. Some examples are shown in Fig. 3, corresponding to the channels of Fig. 1 and 2.

Theorem: The zero error capacity C_0 of a discrete memoryless channel is bounded by the inequalities

$$-\log \max_{P_i} \sum_{ij} A_{ij} P_i P_j \leq C_0 \leq \min_{P_i(j)} C$$

No counterexample to show the inequality ever occurs

where C is the capacity of any channel with transition probabilities $p_i(j)$ and having the adjacency matrix A_{ij} . The upper bound is fairly obvious. The zero error capacity is certainly less than or equal to the ordinary capacity for any channel since the former requires codes with zero probability of error while the latter requires codes approaching zero probability of error. By minimizing the capacity through variation of the $p_i(j)$ we find the lowest upper bound available through this argument. Since the capacity is a continuous function of the $p_i(j)$ in the closed region defined by $p_i(j) \leq 1$, $\sum_j p_i(j) = 1$, we may write min instead of greatest lower bound.

It is worth noting that it is only necessary to consider a particular channel in performing this minimization, although there are an infinite number with the same adjacency matrix. This one particular channel is obtained as follows from the adjacency matrix. If $A_{ik} = 1$ for a pair ik , define an output letter j with $p_i(j)$ and $p_k(j)$ both differing from zero. Now if there are any three input letters, say $i k l$, all adjacent to each other, define an output letter, say m , with $p_i(m)$ $p_k(m)$ $p_l(m)$ all different from zero. In the graph this corresponds to a complete sub graph with three vertices. Next subsets of four letters or complete subgraphs of four vertices, say $i k l m$, are given an output letter, each being connected to it, and so on. It is evident that any channel with the same adjacency matrix differs from that just described only by variation in the number of output symbols for some of the pairs, triplets, etc., of adjacent input letters. If a channel has more than one output symbol for an adjacent subset of input letters, then its capacity is reduced by identifying these. If a channel contains no element, say for a triplet $i k l$ of adjacent input letters, this will occur as a special case of our canonical channel which has output letter m for this triplet when $p_i(m)$, $p_k(m)$ and $p_l(m)$ all vanish.

The lower bound of the theorem will now be proved. We use the procedure of random codes based on probabilities for the letters P_i , these being chosen to minimize the quadratic form $\sum_{ij} A_{ij} P_i P_j$. Construct

an ensemble of codes each containing M words, each word n letters long. The words in a code are chosen by the following probability method. Each letter of each word is chosen independently of all others and has the value i with probability P_i . We now compute the probability in the ensemble that any particular word is not adjacent to any other word in its code. This probability that the first letter of one word is adjacent to the first letter of a second word is $\sum_{i,j} A_{ij} P_i P_j$, since this sums the cases of adjacency with coefficient 1 and those of non-adjacency with coefficient 0. The probability that two words are adjacent in all letters, and therefore adjacent as words, is $(\sum_{i,j} A_{ij} P_i P_j)^n$. The probability of non-adjacency is therefore $1 - (\sum_{i,j} A_{ij} P_i P_j)^n$. The probability that all $M-1$ other words in a code are not adjacent to a given word is, since they are chosen independently, $[1 - (\sum_{i,j} A_{ij} P_i P_j)^n]^{M-1}$, which is, by a well known inequality, greater than $1 - (M-1)(\sum_{i,j} A_{ij} P_i P_j)^n$, which in turn is greater than $1 - M(\sum_{i,j} A_{ij} P_i P_j)^n$. If we set $M = (1 - \epsilon)^{-1} (\sum_{i,j} A_{ij} P_i P_j)^{-n}$, we then have, by taking ϵ small, a rate as close as desired to $-\log \sum_{i,j} A_{ij} P_i P_j$. Furthermore, once ϵ is chosen, by taking n sufficiently large, we can insure that $M(\sum_{i,j} A_{ij} P_i P_j)^n$ is as small as desired, say, less than δ . The probability in the ensemble of codes of a particular word being adjacent to any other in its own code is now less than δ . This implies that there are codes in the ensemble for which the ratio of the number of such undesired words to the total number in the code is less than or equal to δ . For, if not, the ensemble average would be worse than δ . Select such a code and delete from it the words having this property. We have reduced our rate only by at most $\log(1 - \delta)^{-1}$. Since ϵ and δ were both arbitrarily small, we obtain error-free codes arbitrarily close to the rate $-\log \max_{P_i} \sum_{i,j} A_{ij} P_i P_j$ as stated in the theorem.

For simple channels it is usually more convenient to apply particular tricks in trying to evaluate C_0 instead of the bounds given in this theorem which involve maximizing and minimizing processes. The simplest lower bound, as mentioned before, is obtained by merely finding the logarithm of the maximum number of non-adjacent input letters.

A useful device for establishing an upper bound depends upon the adjacency graph for the input symbols. Suppose two vertices a and b of this graph have the property that they are connected together and every

vertex that a is connected to, b is also connected to (but not necessarily conversely). Then vertex b and all lines connected to b may be eliminated from the graph, leaving an adjacency graph for channels with the same zero error capacity. This may be proved by constructing from any error-free code for channels with the first graph an error-free code with the same number of words for the second graph. This is done by replacing in all words of the first code the letter for vertex b wherever it occurs by the letter for vertex a . This does not change adjacency relations among words since a is adjacent to no points that were not already adjacent to b .

Another device which is useful for finding upper bounds is that of eliminating lines in the graph. Eliminating one or more lines in a graph can only increase or leave constant C_0 , since any zero-error code for the old channel will be zero-error for the new channel. By careful choice of one or more lines to eliminate, the graph may be reduced to one for which C_0 is readily evaluated, and if this C_0 equals the lower bound found by choosing a subset of non-adjacent letters, then this gives the zero-error capacity.

These devices, as well as others, may be described in more general terms with the notion of an adjacency-reducing mapping. Suppose that we can find a mapping of letters into other letters, $i \rightarrow \alpha(i)$, with the property that if i and j are not adjacent in the channel (or graph) then $\alpha(i)$ and $\alpha(j)$ are not adjacent. If we have a zero-error code, then we may apply such a mapping letter by letter to the code and obtain a new code which will also be of the zero-error type, since no adjacencies can be produced by the mapping. If all of the letters i are mapped into a subset of the letters, no two of which are adjacent, then it is easily seen that the zero-error capacity of the original channel is the logarithm of the number of letters in this subset. For, in the first place, by forming all sequences of these letters we obtain a zero-error code at this rate. Secondly, any code in the channel can be mapped into a code using only these letters and containing, therefore, only $2^{C_0 n}$ non-adjacent words.

The capacities, or, more exactly, the equivalent numbers of input symbols for all graphs up to five vertices are shown in Fig. 4. These can all be found readily by the tricks mentioned above, excepting the channel of Fig. 2 mentioned previously, for which we know only that the zero-error

capacity lies in the range $\frac{1}{2} \log 5 \leq C_0 \leq \log \frac{5}{2}$.

All graphs with six vertices have been examined and the capacities of all of these can also be found by these devices, with the exception of four. These four can be given in terms of the capacity of Fig. 2, so that this latter graph is essentially the only unsolved problem up to seven vertices. Graphs with seven vertices have not been completely examined but at least one new situation arises, the analog of Fig. 2 with seven input letters.

Theorem: If two channels have zero-error capacities C_0' and C_0'' , their sum has a zero-error capacity greater than or equal to $\log [\exp(C_0') + \exp(C_0'')]$ and their product a zero-error capacity greater than or equal to $C_0' + C_0''$. If the graph of either of the two channels can be reduced to non-adjacent points by the mapping method, then these inequalities can be replaced by equalities.

Proof: It is clear that in the case of the product, the zero-error capacity is at least $C_0' + C_0''$, since we may form a product code from two codes which are close to C_0' and C_0'' . If these codes are not of the same length, we use for the new code the least common multiple of the individual lengths and form all sequences of the code words of each of the codes up to this length. To prove equality in case one of the graphs, say that for the first channel, can be mapped into non-adjacent points, suppose we have a code for the product channel. The letters for the product code, of course, are ordered pairs of letters corresponding to the original channel. Replace the first letter in each pair in all code words by the letter corresponding to reduction by the mapping method. This reduces or preserves adjacency between words in the code. Now sort the code words into A^n subsets according to the sequences of first letters in the ordered pairs. Each of these subsets can contain at most B^n members, since this is the largest possible number of codes for the second channel of this length. Thus, in total, there are at most $A^n B^n$ words in the code, giving the desired result.

In the case of the sum of the two channels, we first show how, from two given codes for the two channels, to construct a code for the sum channel with equivalent number of letters equal to $A^{1-\delta} + B^{1-\delta}$, where δ is arbitrarily small and A and B are the equivalent number of letters for the two codes. Let the two codes have lengths n_1 and n_2 . The new code will have length n where n is the smallest integer greater than both $\frac{n_1}{\delta}$ and $\frac{n_2}{\delta}$. Now form codes for the first channel and for the second channel for all lengths k from zero to n as follows. Let k equal $an_1 + b$, where a and b are integers and $b < n_1$. We form all sequences of a words from the given code for the first channel and fill in the remaining b letters arbitrarily, say all with the first letter in the code alphabet. We achieve at least $A^{k-\delta n}$ different words of length k none of which is adjacent to

any other. In the same way we form codes for the second channel and achieve $B^{k-n\delta}$ words in this code of length k . We now intermingle the k code for the first channel with the $n-k$ code for the second channel in all $\binom{n}{k}$ possible ways and do this for each value of k . This produces a code n letters long with at least $\sum_{k=0}^n \binom{n}{k} A^{k-n\delta} B^{n-k-n\delta} = (AB)^{-\delta n} (A+B)^n$ different words. It is readily seen that none of these different words are adjacent. The rate is at least $\log(A+B) - \delta \log AB$, and since δ was arbitrarily small, we can achieve a rate arbitrarily close to $\log(A+B)$.

To show that it is not possible, when one of the graphs reduces to non-adjacent points, to exceed the rate corresponding to the number of letters $A+B$, consider any particular code of length n for the sum channel. The words in this consist of sequences of letters each corresponding to one or the other of the two channels. The words may be subdivided into classes corresponding to the pattern of the choices of letters between the two channels. There are 2^n such classes with $\binom{n}{k}$ classes in which exactly k of the letters are from the first channel and $n-k$ from the second. Consider now a particular class of words of this type. Replace the letters from the first channel alphabet by the corresponding non-adjacent letters. This does not harm the adjacency relations between words in the code. Now, as in the product case, partition the code words according to the sequence of letters involved from the first channel. This produces at most A^k subsets. Each of these subsets contains at most B^{n-k} members, since this is the greatest possible number of non-adjacent words for the second channel of length $n-k$. In total, then, summing over all values of k and taking account of the $\binom{n}{k}$ classes for each k , there are at most $\sum_k \binom{n}{k} A^k B^{n-k} = (A+B)^n$ words in the code for the sum channel. This proves the desired result. We conjecture but have not been able to prove that the equality of this theorem holds in general, not merely under the conditions given.

Theorem: In any code of length n and rate $R > C_0$, $C_0 > 0$, the probability of error P_e will satisfy

$$P_e \geq \frac{1}{2} (1 - e^{\frac{1}{2} - n(R - C_0)}) p_{\min}^n$$

where p_{\min} is the minimum non-vanishing $p_i(j)$. Thus for $R > C_0$, $E(R) \leq -\log p_{\min}$.

Proof: By definition of C_0 there are not more than e^{nC_0} non-adjacent words of length n . With $R > C_0$, among e^{nR} words there must, therefore, be an adjacent pair. The adjacent pair has a common output word which either can cause with a probability at least p_{\min}^n . This output word cannot be decoded into both inputs. At least one, therefore, must cause an error when it leads to this output word. This gives a contribution at least $e^{-nR} p_{\min}^n$ to the probability of error P_e . Now omit this word from consideration and apply the same argument to the remaining $e^{nR} - 1$ words of the code. This will give another adjacent pair and another contribution of error of at least $e^{-nR} p_{\min}^n$. The process may be continued until the number of code points remaining is just e^{nC_0} . At this time, the probability of error must be at least $(e^{nR} - e^{nC_0})e^{-nR} p_{\min}^n$ or the expression given in the theorem.

THEOREM If $N(A+B) \geq N(A) + N(B)$ holds for sum, it holds for product.

Proof: Let $N = 2^{C_0}$, A, B two networks such that $N(A) + N(B) = N(A+B)$

$$\text{Then } N^2(A+B) = N[(A+B)^2] = N(A^2 + AB + BA + B^2)$$

$$\geq N(A^2) + 2N(AB) + N(B^2)$$

$$= [N(A)]^2 + 2N(AB) + [N(B)]^2$$

$$(N(A) + N(B))^2 =$$

$$(N(A))^2 + 2N(A)N(B) + (N(B))^2 \geq (N(A))^2 + 2N(AB) + (N(B))^2$$

$$N(A)N(B) \geq N(AB)$$

But by Theorem $N(A)N(B) \leq N(AB)$,

$$\therefore N(AB) = N(A)N(B)$$



$$\sqrt{5} \leq N_0 \leq \frac{5}{2}$$

Lower Bound for P_{ef} for a Completely Connected Channel with Feedback

Theorem: $P_{ef} \geq (1 - \frac{1}{M}) p_{min}^n$ where M is the number of messages, the channel is assumed completely connected, and p_{min} is the minimum transition probability. Note if $M \geq 2$, $P_{ef} \geq 1/2 p_{min}^n$.

Proof: Choose any two messages m and m' . (If there is only one message, the theorem is trivially true.) Let x_1 and x_1' be the first transmitted letters for m and m' . Since the channel is completely connected, x_1 and x_1' have a common output letter, say y_1 . Determine the second transmitted letters for m and m' if y_1 is received and let these be x_2 and x_2' . These must have a possible common received letter y_2 . Find the third transmitted letters for m when $y_1 y_2$ was received and for m' when $y_1 y_2$ was received. Let these be x_3 and x_3' . Continue this process to give a received sequence y_1, y_2, \dots, y_n which might occur with either m or m' . Each could cause this sequence with probability greater than or equal to p_{min}^n . At the receiver this sequence must be decoded in an unique way, hence one, at least, of m and m' would be decoded incorrectly if it caused this received sequence. Say this is m -- then m can cause errors to the amount at least $\frac{1}{M} p_{min}^n$. Now, eliminating m from further consideration, take any pair of messages from the remaining $M - 1$ (including m'). The same argument may be applied to this pair to give a second source of error, disjoint to the first, to the amount $\frac{1}{M} p_{min}^n$. Continuing in this way, we can arrive at $M - 1$ disjoint sources of error, each at least $\frac{1}{M} p_{min}^n$, a total of at least $(1 - \frac{1}{M}) p_{min}^n$, proving the theorem.

A Lower Bound for P_e when $R > C$

Theorem: For any code with rate $R > C$, $R - C = \delta$, with block length $n > \frac{2 \log 2}{\delta}$, we have

$$P_e \geq \frac{\delta}{4 (R - \log p_{min})}$$

Hence for any fixed $\delta > 0$, P_e is bounded away from zero.

Proof:

$$\begin{aligned} P_e &\geq 1/2 \rho(R - \frac{1}{n} \log 2) \\ &\geq 1/2 \rho(R - \frac{\delta}{2}) \\ &= 1/2 \rho(C + \frac{\delta}{2}) \end{aligned}$$

For any pair of code words (u, v) such that $p(u, v) > 0$, the mutual information $I_{(u, v)}$ satisfies

$$\begin{aligned} \frac{1}{n} I_{(u, v)} &= \frac{1}{n} \log \frac{p_u(v)}{p(v)} \geq \log p_{\min} - \frac{1}{n} \log p(v) \\ &\geq \log p_{\min} \end{aligned}$$

Now whatever distribution $p(u)$ is used, the mean of $\frac{1}{n} I_{(u, v)}$ is less than or equal to C (by the very definition of C). Thus we have a distribution function $\rho(I)$ which is zero for $I < \log p_{\min}$ and whose mean is less than or equal to C . This implies a lower bound on $\rho(C + \frac{\delta}{2})$. In fact, we must have $\rho(C + \frac{\delta}{2})$ greater than or equal to $\frac{\delta/2}{C + \delta/2 - \log p_{\min}}$, for if not, the mean of the

$$\begin{aligned} \text{distribution would be greater than } &\rho(C + \delta/2) \log p_{\min} \\ + [C + \delta/2] [1 - \rho(C + \delta/2)] &= \frac{\delta/2}{C + \delta/2 - \log p_{\min}} \log p_{\min} \\ + (C + \delta/2) \frac{C - \log p_{\min}}{C + \delta/2 - \log p_{\min}} &= C. \end{aligned}$$

This is a contradiction and consequently $P_e \geq 1/4 \frac{\delta}{C + \delta/2 - \log p_{\min}}$

$$\geq 1/4 \frac{\delta}{R - \log p_{\min}}$$

A Lower Bound for P_e

We will say that the input letters in a channel are uniform if each of these letters has the same set of values for transition probabilities to output letters (not necessarily to the same output letters). In the $p_1(j)$ matrix each row is some rearrangement of the numbers in the first row. If this is true, it is clear that the transition probabilities for words of length n in this channel will have the same property. In fact, the transition probabilities from a particular input word will consist of the r^n products that can be formed from the r transition probabilities for the original channel taken n at a time with repetition allowed. Suppose that when these r^n transition probabilities are arranged in order of decreasing value that the total probability after element number d is $Q = Q_n(d)$.

Theorem: In a channel with uniform input letters, r output letters and the function $Q_n(d)$, any block code of length n and rate R has a probability of error P_e satisfying

$$P_e \geq Q_n \left(\left[e^{n(\log r - R)} + 1 \right] \right) .$$

where the brackets denote the integer part.

Proof: Suppose we have given a code with e^{Rn} words. The probability of not making an error, $1 - P_e$, may be computed by taking the probability of use for each word, e^{-Rn} , and multiplying by the sum of the transition probabilities from that word to all output words which are decoded as the given word. When summed over all input words in the code, this gives $1 - P_e$. Thinking in terms of the matrix of word transition probabilities, this means that a certain selected set of entries from each row is added together and the final result multiplied by e^{-Rn} . The total number of entries added in all the different rows is exactly equal to r^n since this is the total number of output words and each is decoded into exactly one input word. The sum of elements in a particular row is increased or unchanged if we take, in place of the given elements, the same number of elements chosen in order of decreasing value. Because of the assumption of uniform inputs, all the rows have the same sequence of values when arranged in monotone decreasing order. Thus our first operation has served to give us the sum of e^{Rn} (one for each row) beginnings of this sequence of various lengths.

If any two of the rows have different numbers of elements added into the sum, we can again increase or leave unchanged the total by equalizing (as nearly as possible) the number of terms from the two rows, since this replaces smaller valued terms by larger ones. Proceeding in this manner we increase or leave constant the sum while holding the total number of terms at exactly r^n . When the equalization of number of entries from rows has proceeded as far as possible, the number in each row will be within one of r^n/e^{Rn} . More precisely, let r^n/e^{Rn} equal $A + B/e^{Rn}$ where A and B are integers and $B < e^{Rn}$. Then B of the rows will have $A+1$ terms and the remaining $e^{Rn} - B$ will have A terms. We will then have

$$1 - P_e \leq e^{-Rn} [B(1 - Q(A+1)) + (e^{Rn} - B)(1 - Q(A))]]$$

$$P_e \geq e^{-Rn} BQ(A+1) + (1 - e^{-Rn} B)Q(A)$$

$$\geq Q(A+1)$$

(See next page)

$$= Q\left(\left\lceil e^{n(\log r - R)} + 1 \right\rceil\right)$$

Lower Bound with One Type of Input and Many Types of Output

The inequality we have proved holds in any case where the inputs are uniform. However, it may be strengthened in certain cases. Suppose that the input letters (or words) are uniform in the sense previously defined and that the output letters (or words) can be partitioned into a number of subsets S_1, S_2, \dots, S_d with the following property. Each input letter (or word) has the same set of transition probabilities leading to words in S_i as any other input word, for each i . Thus, the channel looks uniform for output words when only the input words and output words in any particular S_i are considered. Let N_i be the number of output words in S_i . Let $Q_i(d)$ be the probability in the tail for S_i analogous to the $Q(d)$ of the preceding theorem. Thus $Q_i(d)$ is the total probability after d elements in the monotone decreasing ordered sequence of all probabilities from an input word to the output words in S_i .

We may argue precisely as we did before for each particular S_i and obtain a lower bound for the probability of errors occurring with received signals in the set S_i . The total probability of error P_e is greater than or equal to the sum on i of these individual contributions.

$$P_e \geq \sum_i Q_i(n_i e^{-nR} + 1) \quad .$$

A more general case may be defined as follows. Suppose the input words can be partitioned into subsets T_1, T_2, \dots, T_c and the output words into subsets S_1, S_2, \dots, S_d , and the channel is uniform in transitions from input set T_i to output set S_j , that is, every member of T_i has the same array of transition probabilities to members of S_j . It is always trivially possible to perform such a partitioning by placing all input letters in different subsets and all output letters also in different subsets. More significantly, if we consider words of length n , we may perform this partitioning by subdividing the input words into subsets according to their composition in terms of letters. Thus, if the letters in the channel are a, b, \dots, g , a composition is defined by a set of integers n_a, n_b, \dots, n_g whose sum is n . All words with exactly n_a a 's, n_b b 's, \dots , n_g g 's will be placed in the corresponding input class. This class would then have $n! / n_a! n_b! \dots n_g!$ members. In an exactly similar way the output words of length n can be partitioned into composi-

tions in terms of output letters. It is immediately seen that each word in a particular input class has the same transition probabilities to a certain output class as any other word in the same input class. Thus this decomposition is of the type we are considering.

We return now to the calculation of a lower bound for P_e , with a given number of code words $M = e^{Rn}$. Our procedure is similar to that used previously; we perform operations which reduce (or leave unchanged) the probability of error and arrive eventually at an easily computed value. Suppose a given code has M_i members in input class T_i ($i = 1, 2, \dots, c$). Let N_j as before be the total number of words in output class S_j and let $Q_{ij}(d)$ be the total probability in the tail beyond entry d when the transition probabilities from a member of set T_i to the words in S_j are arranged in a monotone decreasing sequence. There will be errors in output set S_j at least to the amount $\min_i Q_{ij}(N_j e^{-Rn} + 1)$. This is true since we may reduce the probability of error by equalizing the tails as before for all words from the same input class. Then one may again reduce or leave unchanged the probability of error by replacing words from other input classes by that which minimizes the expression. The details are simple. The total P_e can be bounded from below by summing this over-all output class:

$$P_e \geq \sum_j \min_i Q_{ij}(N_j e^{-Rn} + 1).$$

Another lower bound can be obtained by a slightly different argument. If there are c input classes and e^{Rn} input words, there must be a class with at least e^{Rn}/c input words. If class i contains this many words, the probability of error will be bounded at least by $P_e \geq \sum_j Q_{ij}(N_j c e^{-Rn} + 1)$ since the situation is that covered by the uniform input result. If we minimize this on i , then we will certainly have a lower bound for P_e regardless of which class contains the e^{Rn}/c or more code words. Thus

$$P_e \geq \min_i \sum_j Q_{ij}(N_j c e^{-Rn} + 1).$$

A somewhat stronger but more complex lower bound on P_e can be obtained by a still different variation of these arguments.

Let

$Q_{ij}(P)$ = probability in the tail of the monotone sequence of transition probabilities from input set i to output set j , the tail consisting of probabilities less than P in value.

$N_{ij}(P)$ = total number of terms in this sequence with probabilities greater than or equal to P .

N_j = total number of words in output set j .

Then we will show that the probability of error P_e satisfies

$$P_e \geq \min_i \sum_j Q_{ij}(P_j) \quad (1)$$

~~$\sum_j Q_{ij}(P_j)$~~ $\sum \alpha_i = 1$

Where the P_j satisfy

$$N_j \leq M \sum_i \alpha_i N_{ij}(P_j) \quad (2)$$

The argument here is similar to those before. We assume $\alpha_i M$ messages coded into input set i . To obtain the maximum probability in the parts of the tails of the distributions they should be equalized as nearly as possible to end at the same value of probability for the last term taken.

While this equalization will not, in general, come out even, a value P_j satisfying (2) will be small enough that all the tails of the different sequences beyond this P_j will cause error after the nearest possible equalization. Thus, P_e will have a lower bound given by (1). The minimizing, of course, takes account of the most favorable possible way of dividing the M messages among the input classes.

Application of "Sphere-packing" Bounds to Feedback Case.

In the uniform input case, the lower bounds on the probability of error based on the sphere-packing type of argument apply also to memoryless discrete channels which have a feedback link giving information at the transmitter concerning the previous received letter.

To show this, suppose we have such a uniform input case where the input letters all have the same set of transition probabilities going to output letters. Suppose we have a block code for the feedback system of length n . This means that at the transmitting point there is a device with two inputs, or, mathematically, a function with two arguments. One argument is the message to be transmitted, the other, the past received letters (which have come in over the feedback link). The value of the function is the next letter to be transmitted. Thus, the function may be thought of as $x_{j+1} = f(k, v_j)$ where x_{j+1} is the $j+1$ transmitted letter in a block, k is an index ranging from one to e^{Rn} , and represents the specific message, and v_j is a received word of length j . Thus j ranges from 0 to $n-1$ and v_j over all received words of these lengths.

In operation, if message m_k is to be sent f is evaluated for $f(k, \text{---})$ where the --- means "no word" and this is sent as the first transmitted letter. If the feedback link sends back α , say, as the received letter, the next transmitted letter will be $f(k, \alpha)$. If this is received as β , the next transmitted letter will be $f(k, \alpha, \beta)$, etc.

Remembering our assumption about uniformity, the first transmitted letter for any message gives rise to a set of received letters with probabilities q_1, q_2, \dots, q_t (these being the transition probabilities from any letter).

In each case, (that is, each (message, received letter) pair), a second transmitted letter is determined by the function f . Since the letters are uniform, each gives rise to a second set of letters with probabilities q_1, q_2, \dots, q_t . The probabilities are the same in all cases although the letters to which they apply may differ. Thus, for each message choice m_k there exists a set of possible received two-letter sequences with the same set of probabilities, namely, all pairs $q_i q_j$. Continuing in this manner, each message m_k when fully transmitted gives rise at the receiver to a set of possible received words of length n with the same array of probabilities (regardless of the particular message or the particular noise). These probabilities are the set of all n^{th} degree products of terms from q_1, q_2, \dots, q_t .

At the receiver, a received word must be decoded in an unique way. The probability of error when message m_1 is transmitted is the sum of the above-mentioned transition probabilities to all words of length n which are not decoded as m_1 . If a_1 received words are decoded as message m_1 , then $\sum_1^{a_1} a_i = N$, the total number of different received words of length n . If the transition probabilities are arranged in monotone decreasing order, the probability of errors for message m_k is greater than or equal to the sum of terms in this decreasing sequence after term a_1 , since the sum of the first a_1 terms of a monotone decreasing sequence overbounds the sum of any other a_1 terms. Thus, our estimate of P_e is decreased by taking the first a_1 terms for each message m_1 .

Since the sequences for the different messages m_1 are actually the same, it is again decreased by equalizing, as nearly as possible, the different a_1 . This gives the simplest lower bound on P_e .

The more involved and sharper result, where the different classes of received words are considered, follows by essentially the same argument, on noticing that each transmitter choice gives rise to the different q_{12} transition probabilities and that the equalization may be carried out within these classes as before, always reducing the estimate of P_e .

While it seems likely that the more general results, where the input letters are not uniform, (or slight modifications of these results) hold for the feedback case, no proof has been found. There is, indeed, some extra difficulty here because the transmitter can now take positive and useful action depending on the results at the receiver of earlier parts of the message. In the uniform input case, no very significant action is possible, since all the letters are statistically alike so far as the sphere-packing properties are concerned.

Theorem: Suppose in a channel words of length d can be partitioned into $e^{C_1 d}$ completely connected subsets and we have given a code of length $n+d$ with M words and with probability of error P_{eL} . Then we can construct a code of length n with at least $\frac{1}{2} M e^{-C_1 d}$ words and with probability of error $P_{es} \leq 2 p_{\min}^{-d} P_{eL}$, where p_{\min} is the smallest (nonvanishing) $p_i(j)$ for the channel. If $C_0 = 0$ for the channel we can construct, more strongly, the code of length n with M words and probability of error $P_{es} \leq p_{\min}^{-d} P_{eL}$.

Corollary: Let (R_1, E_1) be any point on the reliability curve for a channel. Construct the straight line through this point and the point $(C_1, \log p_{\min}^{-1})$. The reliability curve lies below or on this straight line for $C_1 < R < R_1$ and above or on it for $R > R_1$. In particular, it lies below the line segment joining $(C_1, \log p_{\min}^{-1})$ and $(C, 0)$ where C is the capacity of the channel.

Proof: We will refer to the given code of length $n+d$ as the long code and codes of length n derived from it as short codes. For simplicity, we will first consider the case where $C_0 = 0$. The short code is then obtained by merely deleting the last d letters of each of the words in the long code. Thus, in the long code, let us designate the words by $T_1 + U_1, T_2 + U_2, \dots, T_M + U_M$.

These words correspond to the M different messages and some of them may consist of the same sequence of input letters (although in general for a good code, this would not be the case).

The short code consists of the words T_1, T_2, \dots, T_K . The decoding process for the short code will be maximum likelihood. Thus, if the received words corresponding to the short code are V_1, V_2, \dots , and V_j is received, it is decoded as that T_i with maximum conditional probability given V_j . Since the T_i are used with equal probability, this is the T_i whose probability of causing V_j is a maximum. We now show that the probability of error in the short code when a particular V , say V_j , is received is less than or equal to p_{\min}^{-d} multiplied by the corresponding probability for the long code, where we must, of course, consider all the possible received signals W_1, W_2, \dots , corresponding to the U part of the long code. Let T_{ML} be the maximum likelihood detection for the short code when V_j is received, and let U^* be the U part of the long code for T_{ML} . Also let the long code decoding system, when V_j and U^* is received, decode it as the message $\phi(k)$, that is, decide that $T_{\phi(k)} = U_{\phi(k)}$ was transmitted.

Since $C_0 = 0$, each pair of words of length d have a possible common received word. In particular, for each k , $U_{\phi(k)}$ and U_{ML} have a W in common. Hence we can find a set of W 's, $W_\alpha, W_\beta, W_\gamma, \dots$, such that each one is a possible result of U_{ML} and one or more of the $U_{\phi(k)}$, and every $U_{\phi(k)}$ has some W in the subset as a possible result. Now the error in the short code (when V_j is received) is given by $P_{es, V_j} = \sum_{i \neq ML} P_{V_j}(U_i)$, that is, the probability of all other transmitted words except the maximum likelihood one (conditional on the received V_j). Consider the probability of error for the long code when V_j is received and in particular those errors resulting when U_i or U_{ML} is received as W_α (say), W_α being the W common to U_i and U_{ML} . Either of U_i or U_{ML} can cause W_α with probability greater than or equal to p_{\min}^{-d} . Whether W_α is decoded as i or ML (or some other way), errors will occur with probability $\geq p_{\min}^{-d} P_{V_j}(U_i)$ since $P_{V_j}(U_{ML}) \geq P_{V_j}(U_i)$, (since U_{ML} was the maximum likelihood U). If there are several U_i 's leading into W_α , we will again have errors caused with probability at least $p_{\min}^{-d} \sum P_{V_j}(U_i)$, summed over this set of i , since if W_α is decoded as one of the U_i , the larger $P_{V_j}(U_{ML})$ takes its place. In total, then, summing over all the W 's in our selected subset, we get a total probability of error for the long code, when V_j is received, $p_{\min}^{-d} \sum_{i \neq ML} P_{V_j}(U_i) \geq p_{\min}^{-d} P_{es, V_j}$. Summing this inequality over all V_j with appropriate probabilities for the V_j , we obtain the desired result

$$P_{es} \leq p_{\min}^{-d} P_{eL}.$$

Now consider the case when $C_1 > 0$. We can subdivide the set of U_1 into $e^{C_1 d}$ subsets such that any two U_1 in the same subset are adjacent. This subdivision partitions the M code words for the long code into $e^{C_1 d}$ subsets, giving $e^{C_1 d}$ codes for each of which the preceding argument will apply. For each of these, therefore, the probability of error for the short code is less than or equal to p_{\min}^{-d} multiplied by the probability of error for the corresponding part of the long code. By the combinatorial argument used in connection with previous results, at least half the code words are in codes of at least half average size, and the average error for these code words is not greater than $2p_{\min}^{-d} P_{eL}$. Hence, there exists among these a code containing at least $\frac{1}{2} M e^{-C_1 d}$ words and with probability of error $P_{es} \leq 2 p_{\min}^{-d} P_{eL}$.

To prove the corollary, let the rate and the reliability of the given long code be R_1 and E_1 , s:

$$R_1 = \frac{1}{n+d} \log M$$

$$E_1 = \frac{1}{n+d} \log P_{eL}^{-1}$$

Further, let the rate and reliability of the short code constructed from this be R and E .

$$R \geq \frac{1}{n} \log M - \frac{1}{n} C_1 d - \frac{1}{n} \log 2 = (1+x)R_1 - xC_1 - \frac{1}{n} \log 2$$

$$E \geq \frac{1}{n} \log P_{eL}^{-1} + \frac{d}{n} \log p_{\min} + \frac{1}{n} \log 2 = (1+x)E_1 + x \log p_{\min} + \frac{1}{n} \log 2$$

where $x = \frac{d}{n}$. If now we consider a series of codes with increasing n approaching the E_1 and R_1 of a point on the curve, then the last terms above approach zero and the E and R of the corresponding series of short codes have a limit supremum on or above the straight line defined by the equations.

$$R_2 = (1+x)R_1 - xC_1$$

$$E_2 = (1+x)E_1 + x \log p_{\min}$$

This straight line passes through the point $(C_1, \log p_{\min}^{-1})$ and the point (R_1, E_1) . The range $x > 0$ for which our statement is true corresponds to points to the right of (R_1, E_1) on the straight line. To the left of

(R_1, E_1) the reliability curve must be on or below this straight line, for if it were above the line, say at (R_3, E_3) , we could use this point for the (R_1, E_1) and obtain a higher value by the construction of these short codes at the original R_1 rate.

This result, it may be noted, is very similar to Theorem . Taken together, they allow one to pass two straight lines through any given point on the reliability curve, and assert that the curve lies within one acute angle to the left of the given point and within the opposite acute angle to the right.

A consequence of this construction is that E , regarded as a function of R , is continuous at least for $C_1 < R < C$ and also that R , regarded as a function of E , is continuous at least for $0 < R < P(C_1)$. This is evident since, for any point inside these intervals, the straight line upper and lower bounds force E (or R) to approach the given point as R (or E) does so.

Theorem: If we have a code with M words, each of length n and with probability of error P_e , we can construct a code of at least $\frac{1}{2} A^{-d} M$ words of length $n-d$ and with a probability of error $P'_e \leq 2 P_e$, where A is the number of distinct input letters.

Proof: Subdivide the M given words into A^d subsets according to the first d letters. The first subset consists of all the code words containing the first input letter in all of the first d positions. The second subset contains the first letter in the first $d-1$ positions and the second letter in its d^{th} position and so on, lexicographically.

At least half of the original words must be in subsets with $\frac{1}{2} A^{-d} M$ or more members, for the total number of words in not more than A^d subsets each of size not more than $\frac{1}{2} A^{-d} M$ is less than or equal to $\frac{1}{2} M$, that is, less than half the total. Hence the other half is in larger subsets. Now consider these larger subsets. The average probability of error in the original code for all words in these subsets is less than or equal to $2 P_e$, since, if not, the average probability of error for all words would be greater than P_e . The probability of error for these larger subsets is a weighted average of the probabilities of error for the individual larger subsets; hence, there exists an individual subset with a probability of error less than or equal to $2 P_e$. If these words alone are used, the probability of error can only be improved, and if the first d letters are deleted, the probability of error is unchanged.

If $d = kn$, $E = -\frac{1}{n} \log P_e$ and $R = \frac{1}{n} \log M$. Then we find for the new code, as $n \rightarrow \infty$,

$$R_1 \rightarrow \frac{1}{1-k} (R - k \log A)$$

$$E_1 \rightarrow \frac{1}{1-k} E$$

This means that on the E, R plot, if a straight line be passed through the curve at E, R and through the point $E = 0, R = \log A$, then the E, R curve lies below (or on) the straight line to the right of the given point and above (or on) the straight line to the left of the given point.

12

A Result for the Memoryless Feedback Channel.

Theorem: Given a code for a memoryless feedback channel, with block length n , probability of error P_e , and number of messages M , we can find a code with block length $n-d$, probability of error $\leq 2 P_e$ and number of messages $\geq M/(ab p_{\max})^d$, where a is the number of letters in the input alphabet, b that for the output alphabet, p_{\max} the largest transition probability and d any desired integer from 0 to n .

Proof: For the given code consider the set of transmission "starts" of length d . Input letter x_1 say is received as y_1 , next x_2 is transmitted and received as y_2 , etc. to x_{d-1} as y_{d-1} and finally x_d as y_d . There are $(ab)^d$ possible starts $(x_1, y_1, x_2, y_2, \dots, x_d, y_d)$ of length d . In the given code let these occur with probabilities q_1, q_2, \dots, q_T (where $T = (ab)^d$). Let the final probability of error for each of these be P_{ei} ($i = 1, 2, \dots, T$). Then $\sum q_i P_{ei} = P_e$. Using our combinational lemma there is at least one of these starts, α , with a $q_\alpha \geq \frac{1}{T} = 1/2T$ and with a $P_{e\alpha} \leq 2 P_e$. Any message which can cause a particular start (such as start α) leads to this start with probability $g = p_{x_1}(y_1) p_{x_2}(y_2) \dots p_{x_d}(y_d)$ where the x_i and y_i are those for the start. The total probability of the start is then $1/M$ times the number of messages that can cause the start times this product. For start α this total probability is $\geq 1/2T$, hence the number of messages must be greater than $1/2T / 1/M g \geq M/2T p_{\max}^d = M/2(ab p_{\max})^d$.

The code to be used of length $n-d$ consists of the messages in the group α , sending only the last $n-d$ letters as though they had started in the manner leading to start α . We have seen that the number available is as stated in the theorem. If the detection system used is that for the original code and all received signals not decoded as one of the messages in the group is counted as an error, then the probability of error will be exactly $P_{e\alpha} \leq 2 P_e$. A suitable distribution of these other received words can only improve this value.

Continuity of $P_{e \text{ opt}}$ as a function of transition probabilities.

Theorem: The probability of error for the optimal code of length n in the channel defined by $p_i(j)$, that is, $P_{e \text{ opt}}(p_i(j), n)$, is a continuous function of $p_i(j)$ in the region R defined by $\sum_j p_i(j) = 1 (i=1, 2, \dots, a)$.

Proof: For a given finite number of input words and a finite number of output words there are a finite number of codes containing M words. There is also only a finite number of decoding systems for each of these codes. Hence there is a finite number of complete systems. Let these be numbered and let the probability of error for the i th one be $P_{ei} (i=1, 2, \dots, f)$. Then

$$P_{e \text{ opt}}(p_i(j), n) = \min_1 P_{ei}(p_i(j), n) \quad (1)$$

Each P_{ei} is a continuous function of $p_i(j)$ in the region R . In fact each P_{ei} is a multinomial in these probabilities, namely, M^{-1} times the sum of the probabilities of each code word being carried to all the received words which are not decoded as the word in question. The minimum of a finite number of continuous functions is a continuous function, proving the theorem. In fact, we may say more strongly that $P_{e \text{ opt}}$ is made up of a finite number of multinomials, each representing $P_{e \text{ opt}}$ in a region of the $p_i(j)$ space.

Codes of a fixed composition.

Consider words of length n . Suppose an input word has $\lambda_i n$ occurrences of the i th letter ($i = 1, 2, \dots, a$). Then we will call the vector λ_i the composition of the word. Similarly, if an output word has $\mu_i n$ occurrences of the i th letter ($i = 1, 2, \dots, b$), then μ_i is the composition of this output word. The number of different compositions of input words is $\binom{n+a}{a} \leq n^a$. The number of different compositions of output words is similarly $\binom{n+b}{b} \leq n^b$. We may consider, for a given channel, codes in which we artificially restrict the input words to a particular composition, say λ_i . We can then consider problems of finding the optimal code and its optimal probability of error and reliability. This reliability we denote by $E(R, \lambda_i, n)$. We will now show the following:

$$\frac{1}{n} \log 2 + \max_{\lambda_i} E(R - \frac{1}{n} \log 2n^a, \lambda_i, n) \geq E(R, n) \geq \max_{\lambda_i} E(R, \lambda_i, n) .$$

The right hand relation is clear since the right hand member is the best reliability for codes all of whose words have the same composition, while $E(R, n)$ is the best reliability with no such restriction and consequently is at least as good. The left hand relation is proved as follows. In a code with e^{Rn} input words distributed over not more than n^a different compositions, the average composition has at least e^{nR}/n^a input words. Using a combinatorial principle previously proved, at least half of the words are in composition classes which contain at least half this average number of words. When a word is in such a class, the probability of error is at least as great as if there were no other input words (except those in the class) and the code was the best possible for the number in the class. The probability of error would again be reduced if the composition in question were that which has the smallest probability of error for the given number of input words. Translating into reliability and rate, the reliability for the cases at hand is not greater than $\max_{\lambda_i} E(R - \frac{1}{n} \log 2n^a, \lambda_i, n)$. Since these words occur at least half the time, the reliability $E(R, n)$ for the original code satisfies the left inequality.

Relation of P_e to ρ

Theorem: Suppose a particular code has e^{nR} words and the distribution function for the information I is $\rho(x)$ (the words being used with equal probability). Then the optimal detection system for this code gives a probability of error P_e satisfying the inequalities

$$\frac{1}{2} \rho(R - \frac{1}{n} \log 2) \leq P_e \leq \rho(R - \frac{1}{n} \log 2)$$

Proof: We first prove the lower bound. By definition of the function ρ , the probability $= \rho(R - \frac{1}{n} \log 2)$ that

$$\frac{1}{n} \log \frac{p(u,v)}{p(u)p(v)} \leq R - \frac{1}{n} \log 2$$

or

$$\frac{p(u,v)}{p(u)p(v)} \leq \frac{1}{2} \cdot e^{nR}$$

or (using the fact that $p(u) = e^{-nR}$)

$$p_v(u) < \frac{1}{2}.$$

Now fix attention on these pairs (u,v) for which this inequality $p_v(u) \leq 1/2$ is true, and imagine the corresponding (u,v) lines to be marked in black and all other (u,v) connecting lines marked in red. We divide the v points into two classes: C_1 consists of those v 's which are decoded into u 's connected by a red line (and also any v 's which are decoded into u 's not connected to the v 's); C_2 consists of v 's which are decoded into u 's connected by a black line. We have established that with probability

$p(R - \frac{1}{n} \log 2)$ the (u, v) pair will be connected by a black line. The v 's involved will fall into the two classes C_1 and C_2 with probability p_1 , say, and $p_2 = p(R - \frac{1}{n} \log 2) = p_1$. Whenever the v is in C_1 an error is produced since the actual u was one connected by a black line and the decoding is along a red line (or to a disconnected u). Thus these cases give rise to a probability p_1 of error. When the v in question is in class C_2 , we have $p_v(u) < 1/2$. This means that with at least an equal probability these v 's can be obtained through other u 's than the one in question. If we sum for these v 's the probabilities of all pairs $p(u, v)$ except that corresponding to the decoding system, then we will have a probability at least $p_2/2$ and all of these cases correspond to incorrect decoding. In total, then, we have a probability of error given by

$$P_e \geq p_1 + p_2 \geq \frac{1}{2} p(R - \frac{1}{n} \log 2).$$

We now prove the upper bound. Consider the decoding system defined as follows. If for any received v there exists a u such that $p_v(u) > \frac{1}{2}$, then the v is decoded into that u . Obviously there cannot be more than one such u for a given v since the sum of these would imply a probability greater than one. If there is no such u for a given v , the decoding is irrelevant to our argument. We may, for example, let such u 's all be decoded into the first word in the input code. The probability of error, with this decoding, is then less than or equal to the probability of all (u, v) pairs for which $p_v(u) \leq \frac{1}{2}$. That is,

$$P_e \leq \sum_S p(u, v) \quad (\text{where } S \text{ is the set of pairs } (u, v) \text{ with } p_v(u) \leq \frac{1}{2}).$$

The condition $p_v(u) \leq \frac{1}{2}$ is equivalent to $\frac{p(u, v)}{p(v)} \leq \frac{1}{2}$, or, again, to $\frac{p(u, v)}{p(u)p(v)} \leq \frac{1}{2} p(u)^{-1} = \frac{1}{2} e^{nR}$. This is equivalent to the condition $\frac{1}{n} \log \frac{p(u, v)}{p(u)p(v)} \leq R - \frac{1}{n} \log 2$. The sum $\sum_S p(u, v)$ where this is true is, by definition, the distribution function of $\frac{1}{n} \log \frac{p(u, v)}{p(u)p(v)}$ evaluated at $R - \frac{1}{n} \log 2$, that is,

$$P_e \leq \sum_S p(u, v) = p(R - \frac{1}{n} \log 2).$$

Theorem: Suppose some $p(u)$ for u words of length n gives rise to a distribution $\rho(I)$. Then given any R and any $\epsilon > 0$ there exists a selection of e^{nR} input words and a decoding system such that if these words are used with equal probability, the probability of error P_e is bounded by

$$P_e \leq \rho(R + \epsilon) + 1/2 e^{-n\epsilon}$$

Proof: For a given R and ϵ consider the pairs (u, v) of input and output words and define the set S to consist of those pairs for which $\log \frac{p(u, v)}{p(u)p(v)} > n(R + \epsilon)$. Thinking of the u 's and v 's as two sets of points with connecting lines between, we can imagine the set of lines corresponding to the set S to be colored red. When the u 's are chosen with probabilities $p(u)$, then the probability that the (u, v) pair will belong to the set S is, by definition of ρ , equal to $1 - \rho(R + \epsilon)$.

Now consider the ensemble of signalling codes obtained in the following manner. The integers $1, 2, 3, \dots, M = e^{nR}$ are associated independently with the different possible input sequences u_1, u_2, \dots, u_B with probabilities $p(u_1), p(u_2), \dots, p(u_B)$. This produces an ensemble of codes each using M (or less) input words. If there are B different input words u_i , there will be exactly B^M different codes in this ensemble corresponding to the B^M different ways we can associate M integers with B input words. These codes have different probabilities. Thus the (highly degenerate) code in which all integers are mapped into input word u_1 has probability $p(u_1)^M$. A code in which d_k of the integers are mapped into u_k has probability $\prod_k p(u_k)^{d_k}$. We will be concerned with an average probability of error for this ensemble of codes. By this we mean the average probability of error when these codes are weighted according to the probabilities we have just defined. We imagine that in using one of these codes each integer is used with probability $1/M$. Note that for some particular selections, several integers may fall on the same input word. This input word is then used with higher probability than the others.

In any particular code of the ensemble, our decoding procedure will be defined as follows. If a received v sequence has no red line coming into it (for this v_i there is no (u, v_i) pair in the set S) then we decode (conventionally) as message 1. If there is exactly one integer mapped into a u connected by a red line to this v_i , we decode as the corresponding integer. If there is more than one such integer, we decode as the smallest such integer.

With any particular code in this ensemble the probabilities of using the different u_i will not, in general, be given by $p(u_i)$. However, if we average over the full ensemble, then each u_i will be used with the probability $p(u_i)$, since integers were mapped into u_i in constructing the ensemble with just this probability. This means that in the ensemble average, a pair (u, v) will also occur with the probability $p(u, v)$.

Now let us compute the average probability of error in this full ensemble of codes. In the ensemble a (u, v) pair will not belong to the set S with the probability $\rho(R + \theta)$. We suppose, pessimistically, that each case of this sort produces an error. The remaining $1 - \rho(R + \theta)$ of the time, the (u, v) pair does belong to the set S and consequently

$$\log \frac{p(u, v)}{p(u)p(v)} > n(R + \theta)$$

$$p_v(u) > p(u) e^{n(R + \theta)}$$

Fixing v at v_i , say, we now sum this inequality over all u 's such that (u, v_i) belongs to the set S . This subset of u 's we call S_i . Thus we obtain

$$\sum_{u \in S_i} p_{v_i}(u) > e^{n(R + \theta)} \sum_{u \in S_i} p(u)$$

Now the left member is clearly less than or equal to one, it being the conditional probability that v_i was caused by a member of S_i . The sum in the right member we will denote by Q_i . It is the total unconditional probability for all members of S_i , that is, for all u 's connected to v_i by red lines.

Using these we obtain

$$1 > e^{n(R + \epsilon)} Q_1$$

$$Q_1 < e^{-n(R + \epsilon)}$$

Now consider the conditional probability in the ensemble of codes of an error in decoding when v_1 is received and the correct message is connected to this v_1 by a red line. This probability P_{ei} is given by

$$P_{ei} = \frac{\sum_{K=1}^M \binom{M}{K} Q_i^K (1-Q_i)^{M-K} (K-1)}{\sum_{K=1}^M \binom{M}{K} Q_i^K (1-Q_i)^{M-K} K}$$

The reason for this is that conditional on the given information, the probability, in the ensemble of codes, of the result being caused by one with exactly K integers coded into the Q_i subset is

$$\frac{\binom{M}{K} Q_i^K (1-Q_i)^{M-K} K}{\sum_{K=1}^M \binom{M}{K} Q_i^K (1-Q_i)^{M-K} K}$$

In the case of such a code, the probability of error in decoding is $\frac{K-1}{K}$. Multiplying by this and summing on K gives the P_{ei} expression above. This may be evaluated easily by noting that the denominator is the expectation of a binomial while the numerator is this same expectation less $\sum_{K=1}^M \binom{M}{K} Q_i^K (1-Q_i)^{M-K} = 1 - (1-Q_i)^M$. Hence, we have

$$\begin{aligned} P_{ei} &= \frac{MQ_i - 1 + (1-Q_i)^M}{MQ_i} \\ &= \frac{MQ_i - 1 + 1 - MQ_i + \frac{M(M-1)}{2} Q_i^2 - \dots}{MQ_i} \\ &= \frac{1}{2} (M-1) Q_i - \frac{1}{6} (M-1)(M-2) Q_i^2 + \dots \\ &\leq \frac{1}{2} (M-1) Q_i < \frac{1}{2} M Q_i \leq \frac{1}{2} e^{-n\epsilon} \end{aligned}$$

provided $e^{-n\theta} < 1$, since then $MQ_i < 1$ and the alternating binomial expansion is decreasing in absolute value and hence may be overestimated by dropping the terms after $1/2(M-1)Q_i$.

Now since $P_{ei} < 1/2 e^{-n\theta}$ for each i , the probability of error when the (u,v) pair belongs to set S is less than $1/2 e^{-n\theta}$. Hence, the unconditional probability of error over the ensemble of codes satisfies

$$P_e < p(R+\theta) + [1-p(R+\theta)] \frac{1}{2} e^{-n\theta}$$

$$P_e < p(R+\theta) + \frac{1}{2} e^{-n\theta}$$

This being the average probability of error over the ensemble of codes, there must be at least one particular code in the ensemble with a probability of error this low. This proves the theorem. More generally, one may say that at least half the codes in the ensemble have a probability of error less than twice this bound and at least a fraction δ have a probability of error less than $\frac{1}{\delta}$ times this bound.

A bound on P_e for a random code.

Theorem: Given a distribution $p(u)$ for input words of length n which produces the information distribution $p(x)$, then the random ensemble of codes with e^{Rn} words based on $p(u)$ has an average probability of error satisfying

$$P_e \leq e^{Rn} \int_{R_1}^{\infty} e^{-nx} d\rho(x) + \rho(R_1) \\ = n e^{Rn} \int_R^{\infty} \rho(x) e^{-nx} dx$$

Hence there exist particular codes with e^{Rn} members and this probability of error.

Proof: Construct the random ensemble of codes, each code having e^{Rn} members and based on the given input distribution $p(u)$. We wish to calculate a bound for the average probability of error over this ensemble. In the ensemble, pairs (u, v) of transmitted and received words occur with the same probabilities as in the original situation produced by giving the input words probabilities $p(u)$. We calculate the error probability by an integration on the variable x occurring in the information distribution $\rho(x)$. The probability that a (u, v) pair are such as to give an x lying in the interval $x_i < x \leq x_{i+1}$ is $\rho(x_{i+1}) - \rho(x_i)$. For such a (u, v) pair

$$x_i < \frac{1}{n} \log \frac{p(u, v)}{p(u)p(v)} \leq x_{i+1}$$

or

$$p(u) e^{nx_i} < p_v(u) \leq p(u) e^{nx_{i+1}}$$

If we sum the left terms of this inequality over all words u in set S , say, with greater conditional probability, we obtain

$$e^{nx_i} \sum_S p(u) < \sum_S p_v(u) \leq 1 \\ Q = \sum_S p(u) < e^{-nx_i}$$

since the total probability of set S conditional on v cannot exceed 1. Our detection system will be to choose among the possible words in a

particular code when v is received that one for which $p_v(u)$ (in the original probability system) was greatest. A (u, v) pair in the interval x_i, x_{i+1} will be safe in a code in the ensemble if the set s for that v is empty (apart from the particular w which produced the pair). In the ensemble, the probability of error from a (u, v) pair may be calculated as in the simple threshold case. We obtain an upper bound from this argument

$$\frac{1}{2} e^{Rn} P_e < e^{Rn} P_e \leq e^{n(R-x_i)}.$$

We now can overestimate the probability of error by summing the probability of the (u, v) pair being in the interval x_i, x_{i+1} multiplied by the probability of words in the set S for such a case. Note also that our bound for the latter is one for $x = R$; for smaller x 's we use the bound one rather than $e^{n(R-x_i)}$. As the intervals x_i, x_{i+1} approach zero length, our bound approaches the integral form

$$P_e \leq \rho(R) + \int_R^\infty e^{n(R-x)} d\rho(x)$$

Integrating by parts

$$\begin{aligned} P_e &\leq e^{Rn} \left[\rho(x) e^{-xn} \right]_R^\infty + ne^{Rn} \int_R^\infty \rho(x) e^{-nx} dx + \rho(R) \\ &= ne^{Rn} \int_R^\infty \rho(x) e^{-nx} dx \\ &= e^{Rn} \int_\infty^R \rho(x) dx e^{-nx}. \end{aligned}$$

Corollary: Under the conditions of the theorem, suppose a maximum for $x \geq R$ of $\rho(x)e^{-xn}$ occurs at $x = R_m$. Then

$$P_e \leq e^{(R-R_m)n} \rho(R_m) \left[\log e \rho(R_m)^{-1} + n(R_m - R) \right].$$

In particular, if $R_m = R$

$$P_e \leq \rho(R) \log e \rho(R)^{-1}.$$

Proof: Using the second formula in the theorem, we have

$$P_e \leq ne^{Rn} \int_R^\infty e^{-nx} \rho(x) dx.$$

The maximum of the integrand by the conditions of the corollary is $e^{R_m n} \rho(R_m)$. We also have an upper bound for the integrand $e^{-x n}$, since $\rho(x) \leq 1$. These two bounds cross at $x = a$, where a satisfies

$$e^{-a n} = e^{-R_m n} \rho(R_m)$$

$$a = \frac{1}{n} \log \rho(R_m)^{-1} + R_m.$$

Replacing the integrand by $e^{-R_m n} \rho(R_m)$ for $x \leq a$ and by $e^{-x n}$ for $x > a$, we obtain the upper bound for P_e

$$P_e \leq e^{(R - R_m)n} \rho(R_m) \left[\log \rho(R_m)^{-1} + n(R_m - R) \right].$$

Setting $R_m = R$ gives the second bound.

The Feinstein Bound

It is interesting to compare these results with the bound on the probability of error found by Feinstein. Using a different method of proving the coding theorem for a noisy channel, he found the following upper bound for the probability of error:

$$P_e \leq \frac{1}{1-\delta_2} \left[2^{-\frac{n}{2}(C-R-\epsilon_1-\epsilon_2)} + (\delta_1^+)^{1/2} \right]^2 = U$$

in which

n = block length of the code

C = channel capacity

$R = \frac{1}{n} \log$ (number of code words)

δ_1^+ can be taken to be $\text{Prob} \left[|H(X/Y) - \frac{1}{n} \log p(u/v)| \geq \epsilon_1 \right]$

δ_2 can be taken to be $\text{Prob} \left[|H(X) - \frac{1}{n} \log p(u)| \geq \epsilon_2 \right]$

In using the values above for δ_1^+ and δ_2 we are using the most favorable values to give a low bound on P_e . The bound U above may be approximated within a factor of 2 by a somewhat simpler expression as follows:

$$\frac{1}{1-\delta_2} \left[e^{-n(C-R-\epsilon_1-\epsilon_2)} + \delta_1^+ \right] \leq U \leq \frac{2}{1-\delta_2} \left[e^{-n(C-R-\epsilon_1-\epsilon_2)} + \delta_1^+ \right]$$

The left inequality is obtained by squaring the expression for U and dropping the necessarily positive middle term. The right inequality follows from noting that $2AB \leq A^2 + B^2$ so that U is increased by deleting the middle term and doubling the squared terms.

The bound is somewhat simplified in the case where $p(u)$, the probability of input word u to achieve channel capacity, is constant at $2^{-nH(u)}$. We then have $\delta_2 = 0$ and $\epsilon_2 = 0$. This situation occurs, for example, in channels with uniform input letters, as we have seen previously, and in particular in the binary symmetric channel. In these cases, the inequalities simplify to

$$2^{-n(\Delta-\epsilon_1)} + \delta_1^+ \leq U \leq 2(2^{-n(\Delta-\epsilon_1)} + \delta_1^+),$$

where we define $\Delta = C - R$ as the discrepancy between channel capacity and

rate for the code. Note also in this case that

$$\begin{aligned}
 \delta_1^+ &= \text{Prob} \left[\left| H(X/Y) + \frac{1}{n} \log (u/v) \right| \geq \epsilon_1 \right] \\
 &= \text{Prob} \left[\left| H(X/Y) - H(X) - \frac{1}{n} \log (p(u,v)/p(u)p(v)) \right| \geq \epsilon_1 \right] \\
 &= \text{Prob} \left[\left| \frac{1}{n} \log (p(u,v)/p(u)p(v)) \right| \geq C - \epsilon_1 \right] \\
 &= \rho(C - \epsilon_1)
 \end{aligned}$$

where ρ is the distribution function for information that we have used previously. Making the change of variable $\epsilon_1 = \Delta - \theta$, the inequalities for U become

$$\rho(\Delta - \theta) + 2^{-n\theta} \leq U \leq 2 \left[\rho(\Delta - \theta) + 2^{-n\theta} \right].$$

This may be compared with the inequality () found for the random code by the simple threshold method. It will be seen that they are within at worst a factor of 2 of each other. Since the bound () leads in the binary symmetric channel to a reliability bound considerably poorer than the true reliability curve, the same may be said of the Feinstein bound. We have made no approximations in estimating the reliability bound from the inequality obtained by Feinstein. It follows that either the type of code (or, more precisely, the poorest code that can be constructed by his method) is considerable poorer in reliability than the random code or else that the bound () is a relatively poor estimate of the error probability of these codes (that is, that approximations made prior to this formula were sufficiently crude as to cause this difference in the reliability bounds). Which of these is actually the case we have not determined.

Relations Between Reliability and Minimum Word Separation

In this section we prove some results relating probability of error with the minimum separation between words in the code. These results show that when the signalling rate R is very small the reliability is approximately the minimum separation. As a consequence, to obtain a good code for R near zero, the essential feature is to choose a set of code words such that the minimum separation between any pair is as large as possible.

Theorem: For any code with rate R and maximum likelihood detection

$$\Delta_{\min} - R \leq \frac{1}{n} \log P_e \leq \Delta_{\min} + R - \frac{1}{n} \log 2$$

where Δ_{\min} is the minimum separation between words of the code. Hence, for any code sequence with rate approaching zero and maximum likelihood detection, the reliability approaches the minimum separation.

Corollary: $E(0^+) = \lim_{R \rightarrow 0^+} \lim_{n \rightarrow \infty} \max_{\text{codes of rate } R} \Delta_{\min}$

Proof: Let two words at minimum distance be W_1 and W_2 . The probability of error for the code is certainly at least $\frac{2}{M}$ times the probability of error when W_1 or W_2 is used, since $\frac{2}{M}$ of the time one or the other of these will occur. This latter probability is certainly at least what it would be if none of the other words (except W_1 and W_2) were present, and the detection were by maximum likelihood. This last is $e^{-\Delta_{\min} n}$. Thus

$$P_e \geq \frac{2}{M} e^{-\Delta_{\min} n}$$

Taking the logarithm and dividing by n , we obtain the upper bound.

The lower bound is obtained by noting that the probability of error when a particular word is transmitted can be calculated by summing the probabilities of being interpreted as each other word. These terms are

overestimated by taking each other word to be at separation Δ_{\min} and adding these contributions disjunctively. This amounts to adding $M(M-1)/2$ contributions (one for each pair of words) and giving each the value just obtained $2/M e^{-n \Delta_{\min}}$ for the worst pair, thus

$$P_e \leq \frac{2}{M} \frac{M(M-1)}{2} e^{-n \Delta_{\min}} \leq M e^{-n \Delta_{\min}}$$

By taking logarithms and dividing by n we obtain the desired result.

Since for $R \rightarrow 0$ the two bounds converge to Δ_{\min} , the second statement of the theorem is true. The corollary results on combining the theorem with the definition of the reliability function E .

Corollary: Let $\Delta_{\min}^*(R, n)$ be the minimum separation between words in the code of rate R , block length n , which maximizes this minimum distance for a given channel. Then the reliability characteristic $E(R)$ for the channel satisfies

$$\lim_{n \rightarrow \infty} \Delta_{\min}^*(R, n) - R \leq E(R) \leq \lim_{n \rightarrow \infty} \Delta_{\min}^*(R, n) + R$$

Proof: For the right inequality, note that for any sequence of codes of increasing block length n the $\Delta_{\min}(R, n) \leq \Delta_{\min}^*(R, n)$ (since Δ_{\min}^* is the largest possible Δ_{\min} for the given R and n). Hence for sufficiently large n , all Δ_{\min} in the sequence are less than $\lim_{n \rightarrow \infty} \Delta_{\min}^* + \epsilon$ (for any positive ϵ). Now, using the theorem (and noting that $\frac{1}{n} \log 2 \rightarrow 0$), we obtain $E \leq \lim_{n \rightarrow \infty} \Delta_{\min}^* + R + \epsilon$. This being true for any positive ϵ , it is true for $\epsilon = 0$.

The left inequality also follows easily from the theorem. Take a subsequence from the sequence of codes giving Δ_{\min}^* , which actually approaches $\lim_{n \rightarrow \infty} \Delta_{\min}^*$. Applying the lower bound of the previous theorem to this subsequence of codes, we obtain the left inequality above.

Our next result shows that by selecting our codes the R in the upper bounds of these results can be eliminated.

Theorem: Given a code sequence approaching rate R and reliability E , there exists an expurgated sub-sequence approaching the same rate R and reliability E and with $E \leq \lim_{n \rightarrow \infty} \Delta_{\min}(n)$ where $\Delta_{\min}(n)$ is the minimum separation between words in the n th code in the expurgated sub-sequence.

Proof: For any given Δ perform the following operation. Delete, in each code of the given sequence, one of the points which has a nearest neighbor (provided this separation is less than or equal to Δ). Next, delete one of the points in the resulting code which are closest together, and so on up to the point at which no points remain with a separation less than or equal to Δ . This is done for all the codes in the sequence. For each Δ , either there exists an $\epsilon > 0$ for which an infinite sub-sequence of the codes remaining have a fraction at least ϵ of the original points left or such an ϵ does not exist. This divides values of Δ into two Borel classes and gives a minimum division point Δ_0 such that for $\Delta < \Delta_0$ the ϵ exists and for $\Delta > \Delta_0$ it does not.

Choose any small interval $\delta > 0$ and consider the code sequence resulting for $\Delta = \Delta_0 - \delta$. The rate for this sequence is at least $R + \frac{1}{n} \log \epsilon$ and hence approaches R as $n \rightarrow \infty$. Furthermore, almost all points in the codes remaining in the sequence have a neighbor in the interval $\Delta_0 - \delta$ to Δ_0 , by the construction of Δ_0 . Finally, the E for these codes must be the same as the original E since errors due to points retained are only increased at most in the ratio $\frac{1}{\epsilon}$, due to increased usage of these points. This will not affect E . This code sequence is then ideal and close to uniform in nearest

neighbor separation. Almost all points have a nearest neighbor between $\Delta_0 - \delta$ and $\Delta_0 + \delta$ and δ is arbitrarily small. The argument about P_e given in the preceding theorem can now be improved since almost all points have such a near neighbor. Thus we get the inequality without the R term

$$E \leq \Delta_0 - \Delta_{\min} + \delta$$

for any $\delta > 0$, and hence we can obtain a sub-sequence for which $E \leq \lim_{n \rightarrow \infty} \Delta_{\min}^{(n)}$, as stated in the theorem.

Inequalities for Decodable Codes

Consider codes of the following sort. There are a basic letters and s words W_1, W_2, \dots, W_s formed of sequences of the letters. These words have length $\ell_1, \ell_2, \dots, \ell_s$ (not necessarily equal). The code is supposed to be decodable, by which we mean that any finite sequence of letters can be broken down into words in at most one way.

Theorem: For such a decodable code we have

$$\sum a^{-\ell_i} \leq 1 \quad (1)$$

and

$$\sum p_i \ell_i \geq -\sum p_i \log_2 p_i \quad (2)$$

where the p_i are any set of non-negative numbers such that $\sum p_i = 1$.

Proof: The two inequalities are proved in very similar fashion. We prove (2) first. Choose a set of rational numbers q_i whose sum is one and which are close approximations to the p_i , so close that

$$\left| \sum q_i \ell_i - \sum p_i \ell_i \right| < \epsilon \quad (3)$$

and

$$\left| \sum q_i \log q_i^{-1} - \sum p_i \log p_i^{-1} \right| < \epsilon.$$

This is possible for any $\epsilon > 0$, since both $\sum q_i \ell_i$ and $\sum q_i \log q_i^{-1}$ are continuous functions of the q_i in the range of allowed values. Now consider all sequences of words which contain exactly mq_1 occurrences of word W_1 , mq_2 occurrences of word W_2 , etc. Here, m is any multiple of the least common denominator of the q_i . All of these sequences contain exactly m words and are of length exactly $\sum mq_i \ell_i$. The number of these sequences is at least

$$\frac{m!}{mq_1! mq_2! \dots mq_s!} \approx \frac{e^{-\frac{s}{12}}}{\sqrt{2\pi m \prod q_i}} (\pi q_i^{-1})^m$$

This total number of sequences must be less than or equal to $a^{n \sum q_i \ell_i}$, since this is the total number of possible sequences of the length in question and each of the sequences we have constructed must be different for unique decoding. Thus

$$a^{n \sum q_i \ell_i} \geq \frac{e^{-\frac{s}{12}}}{\sqrt{2\pi m \prod q_i}} (\pi q_i^{-1})^m.$$

Taking logarithms to the base a and dividing by m ,

$$\sum q_i l_i \geq -\sum q_i \log_a q_i - \frac{s}{12m} - \frac{1}{m} \log_a \sqrt{2mm!q_i}.$$

Using (3)

$$\sum p_i l_i \geq -\sum p_i \log_a p_i - \frac{s}{12m} - \frac{1}{m} \log_a \sqrt{2mm!q_i} - 2\varepsilon.$$

Since ε is arbitrarily small and m can be arbitrarily large, we must have the desired relation (2):

$$\sum p_i l_i \geq -\sum p_i \log_a p_i.$$

The inequality (1) is proved as follows. Let $p_i = Aa^{-l_i}$ where A is chosen so that $\sum Aa^{-l_i} = 1$. Choose a set of rational q_i summing to one and approximating to the p_i , in the sense that

$$|\sum p_i l_i - \sum q_i l_i| < \varepsilon$$

$$|\sum p_i \log_a p_i^{-1} - \sum q_i \log_a q_i^{-1}| < \varepsilon.$$

Choose an integer m such that the $q_i m$ are all integers and consider sequences containing exactly $q_i m$ occurrences of word W_i . Thus there are m words in each sequence, and their length is $\sum q_i m l_i$. The total number of sequences we construct is less than or equal to the total number available, since the unique decodability makes them all different. Hence,

$$\frac{m!}{q_1 m! q_2 m! \dots q_s m!} \leq a^{m \sum q_i l_i}.$$

Using the lower bound on the multinomial coefficient as before and taking logarithms to the base a , we arrive at

$$\sum q_i l_i \geq -\sum q_i \log_a q_i - \frac{s}{12m} - \frac{1}{m} \log_a \sqrt{2mm!q_i}.$$

Exactly the same argument as before leads to

$$\sum p_i l_i \geq -\sum p_i \log_a p_i$$

or, replacing p_i by its value Aa^{-l_i} ,

$$\sum Aa^{-l_i} l_i \geq -\sum Aa^{-l_i} \log_a Aa^{-l_i} = \sum Aa^{-l_i} l_i - \log_a A \sum Aa^{-l_i}$$

$$0 \geq -\log_a A$$

$$A = \left(\sum a^{-l_i} \right)^{-1} \geq 1$$

This is the desired result (1).

Convexity of Channel Capacity as a Function of Transition Probabilities

Theorem: The channel capacity for transition probabilities $p_i(j)$ is a convex downward function of these probabilities. That is, the capacity C for the probabilities $r_i(j) = \frac{1}{2}(p_i(j) + q_i(j))$ satisfies the inequality

$$C \leq \frac{1}{2} C_1 + \frac{1}{2} C_2$$

where C_1 is the capacity with probabilities $p_i(j)$ and C_2 that with probabilities $q_i(j)$.

Proof: Let the capacity of the $r_i(j)$ channel be achieved by the input probabilities P_i . Now consider the following channel. There are as many inputs as in the given channels but twice as many outputs, a set j and a set j' . Each input has transitions $\frac{1}{2} p_i(j)$ and $\frac{1}{2} q_i(j')$. Thus, this is the channel we would obtain by halving all probabilities in the $p_i(j)$ and the $q_i(j)$ channels and identifying the corresponding inputs but leaving the outputs distinct. We note that if the corresponding outputs are identified, the channel reduces to the $r_i(j)$ channel. We note also that without this identification the channel looks like one which half the time acts like the $p_i(j)$ channel and half the time the $q_i(j)$ channel. An identification of certain outputs always reduces (or leaves equal) rate of transmission. Let this channel be used with probabilities P_i for the input symbols. Then this inequality in rates may be written

$$H(x) - \left(\frac{1}{2} H_{y_1}(x) + \frac{1}{2} H_{y_2}(x) \right) \geq H(x) - H(y) = C$$

where $H_{y_1}(x)$ is the conditional entropy of x when y is in the j group and $H_{y_2}(x)$ that when y is in the j' group. Splitting $H(x)$ into two parts to combine with the $H_{y_1}(x)$ and $H_{y_2}(x)$, we obtain

$$\frac{1}{2} R_1 + \frac{1}{2} R_2 \geq C$$

where R_1 is the rate for the $p_i(j)$ channel when the inputs have probabilities P_i and R_2 is the similar quantity for the $q_i(j)$ channel. These rates, of course, are less, respectively, than C_1 or C_2 , since the capacities are the maximum possible rates. Hence we get

$$\frac{1}{2} C_1 + \frac{1}{2} C_2 \geq C \quad .$$

ASTERNWALL 78965 WCFox
2-365

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

36 d

A Geometric Interpretation of Channel Capacity

The calculations involved in determining the rate R and channel capacity C for a discrete memoryless channel can be given an interesting geometric formulation that leads to some insights into the properties of these quantities.

Let a channel be defined by the matrix $\|p_i(j)\|$ of transition probabilities from input letter i to output letter j ($i = 1, 2, \dots, a$; $j = 1, 2, \dots, b$). We can think of each row of this matrix as defining a vector or a point in a $b - 1$ dimensional simplex (the $b - 1$ dimensional analog of triangle, tetrahedron, etc.). The coordinates of the point sum to one, $\sum_j p_i(j) = 1$, and they are known as barycentric coordinates. They correspond, for example, to the coordinates a chemist uses when he describes an alloy in terms of the fractions of various components and chemists often plot properties of alloys in a simplex of one, two or three dimensions (line segment, triangle, or tetrahedron).

We thus associate a point or vector \underline{A}_i with input i . Its components are equal to the probabilities of various output letters if only this input were used. If all the inputs are used, with probability P_i for input i , the probabilities of the output letters are given by the components of the vector sum

$$\underline{Q} = \sum_i P_i \underline{A}_i.$$

\underline{Q} is a vector or point in the simplex corresponding to the output letter probabilities. Its j th component is $\sum_i P_i p_i(j)$.

Now, for notational convenience, we define the entropy of a point or a vector in a simplex to be that of the barycentric coordinates of the point interpreted as probabilities. Thus we write

$$H(\underline{A}_i) = - \sum_j p_i(j) \log p_i(j) \quad i = 1, 2, \dots, a$$

$$H(\underline{Q}) = - \sum_j \sum_i P_i p_i(j) \log \sum_i P_i p_i(j)$$

= entropy of received distribution.

In this notation, the rate of transmission R for a given set of input probabilities P_i is given by

$$R = H\left(\sum_{i=1}^a P_i \underline{A}_i\right) = \sum_i P_i H(\underline{A}_i)$$

$$= H(\underline{Q}) = \sum_i P_i H(\underline{A}_i)$$

The function $H(\underline{Q})$ where \underline{Q} is a point in the simplex is a convex upward function. For if the components of \underline{Q} are x_i , we have

$$H = - \sum_i x_i \log x_i$$

$$\frac{\partial H}{\partial x_i} = - (1 + \log x_i)$$

$$H_{ij} = \frac{\partial^2 H}{\partial x_i \partial x_j} = \begin{cases} 0 & i \neq j \\ -\frac{1}{x_i} & i = j \end{cases}$$

Hence $\sum_{ij} H_{ij} \Delta x_i \Delta x_j = - \sum_i \frac{1}{x_i} (\Delta x_i)^2$ is a negative definite form. This is true in the space of all non-negative x_i and, hence, certainly in the subspace where $\sum x_i = 1$. It follows that the rate R above is always non-negative and, indeed, since H is strictly convex (no flat regions), that R is positive unless $\sum P_i \underline{A}_i = \underline{A}_s$ whenever $P_s \neq 0$.

The process of calculating R can be visualized readily in the cases of two or three output letters. With these output letters, imagine an equilateral triangle on the floor for the simplex containing the points \underline{A}_i

and Q . Above this triangle is a rounded dome like the Kresge Auditorium. The height of the dome at any point A is $H(A)$. If there were three input letters with corresponding vectors A_1, A_2, A_3 these correspond to three points in the triangle and, straight up from these, to three points on the dome. Any received vector $Q = \sum P_i A_i$ is a point within the triangle on the floor defined by A_1, A_2, A_3 . $H(Q)$ is the height of the dome above the Q point and $\sum P_i H(A_i)$ is the height above Q of the plane defined by the three dome points over A_1, A_2, A_3 . In other words, R is the vertical distance over Q from the dome down to the plane defined by these three points.

The capacity C is the maximum R . Consequently in this particular case it is the maximum vertical distance from the dome to the plane. This clearly occurs at the point of tangency of a plane tangent to the dome and parallel to the plane defined by the input letters.

If there were four input letters, they would define a triangle or a quadrilateral on the floor depending on their positions, and their vertical points in the dome would in general define a tetrahedron. Using them with different probabilities would give any point in the tetrahedron as the subtracted value $\sum P_i H(A_i)$. Clearly, the maximum R would occur by choosing probabilities which place this subtracted part on the lower surface of the tetrahedron.

These remarks also apply if there are still more input letters. If there are a input letters they define an a -gon or less in the floor and the vertically overhead points in the dome produce a polyhedron. Any point in the convex hull of the points obtained in the dome can be reached with suitable choice of the P_i and corresponds to some subtracted term in R .

It is clear that to maximize R and thus obtain C one need only consider the lower surface of this convex hull.

It is also clear geometrically, from the fact that the lower surface of the polyhedron is convex downward and the dome is strictly convex upward, that there is a unique point at which the maximum R , that is C , occurs. For if there were two such points, the point halfway between would be even better since the dome would go up above the line connecting the points at the top and would be at least as low at the bottom surface. The rate R is thus a strictly convex function of the received vector \underline{Q} .

It is also true that the rate R is a convex upward function of the input probability vector (with a barycentric coordinates P_1, P_2, \dots, P_a rather than the b coordinates of our other vectors). This is true since the \underline{Q} vectors \underline{Q} and \underline{Q}' corresponding to the input probabilities P_1 and P'_1 are given by

$$\underline{Q} = \sum_i P_i \underline{A}_i$$

$$\underline{Q}' = \sum_i P'_i \underline{A}_i$$

The \underline{Q} corresponding to $\alpha P_1 + \beta P'_1$ (where $\alpha + \beta = 1$ and both are positive) is $\alpha \underline{Q} + \beta \underline{Q}'$ and consequently the corresponding $R \geq \alpha R + \beta R'$, the desired result. The equality can occur when $\underline{Q} = \underline{Q}'$, so we cannot say in this case a strictly convex function.

These last remarks also imply that the set S of P_1 vectors which maximize the rate at the capacity C form a convex set in its a -dimensional simplex. If the maximum is obtained at two different points it is also attained at all points on the line segment joining these points. Furthermore, any local maximum of R is the absolute maximum C , for if not, join the points corresponding to the local maximum and the absolute maximum. The value of R must lie

on or above this line by the convexity property, but must lie below it when sufficiently close to the local maximum to make it a local maximum. This contradiction proves our statement.

Another property we may deduce is that the capacity C can always be attained using not more than b of the input letters. This is because any point on the surface of a b -dimensional polyhedron is interior to some face. This face may be subdivided into $b - 1$ dimensional simplexes (if it is not already a simplex). The point is then interior to one of these. The vertices of the simplex are b input letters, and the desired point can be expressed in terms of these.

This picture gives considerable information concerning which input letters should be used to achieve channel capacity. If the vector \underline{A}_t , say, corresponding to input letter t , is interior to the convex hull of the remaining letters, it need not be used. Thus, suppose $\underline{A}_t = \sum_{i \neq t} \alpha_i H(\underline{A}_i)$ where $\sum_i \alpha_i = 1, \alpha_i \geq 0$. Then by the convexity properties $H(\underline{A}_t) \geq \sum_{i \neq t} \alpha_i H(\underline{A}_i)$. If by using the \underline{A}_i with probabilities P_i we obtain a rate $R = H(\sum P_i \underline{A}_i) - \sum P_i H(\underline{A}_i)$, then a rate greater than or equal to R can be obtained by expressing \underline{A}_t in terms of the other \underline{A}_i , for this leaves unaltered the first term of R and decreases or leaves constant the sum.

In the case of only two output letters the situation is extremely simple. Whatever the number of input letters, only two of them need be used to achieve channel capacity. These two will be those with the maximum and minimum transition probabilities to one of the output letters. These values, P_1 and P_2 , say, are then located in the one-dimensional simplex, a line segment of unit length, and projected upward to the H -curve as shown in Fig. 1. The

secant line is drawn and the capacity is the largest vertical distance from the secant to the curve. The probabilities to achieve this capacity are in proportion to the distances from this point to the two ends of the secant.

In the case of three output letters, the positions of all vectors corresponding to input letters may be plotted in an equilateral triangle. The circumscribing polygon (convex hull) of these points may now be taken and any points interior to this polygon (including those on edges) may be deleted. What is desired is the lower surface of the polyhedron determined by the points in the H-surface above these points. This lower surface, in general, will consist of triangles and the problem is to determine which vertices are connected by edges. A method of doing this is to consider a line joining a pair of vertices and then to calculate for other lines whose projections on the floor cross this line, whether they are above it or below it in space. If there is no line below the first line, this line is an edge on the lower surface of the polyhedron. If a second line is found below the first line this one may be tested in a similar fashion, and eventually an edge is isolated. This edge divides the projection into two smaller polygons and these may now be studied individually by the same means. Eventually, the original polygon will be divided by edges into a set of polygons corresponding to faces of the polyhedron. Each of these polygons may then be examined to determine whether or not the point of tangency of the parallel plane which is tangent to the H-surface lies over the polyhedron. This will happen in exactly one of the polygons and corresponds to the Q for maximum R .

Log Moment Generating Function for the Square of a Gaussian Variate

Suppose x is a gaussian random variable with variance σ^2 . Its density function is

$$p(x)dx = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} dx.$$

The random variable $u = x^2$ will have a density distribution $q(u)$ obtained by substituting $x = \sqrt{u}$, $dx = du / 2\sqrt{u}$ and then multiplying the result by 2. This last operation takes account of the two halves of the original distribution which both go into the positive u range. The result of these substitutions is

$$q(u) du = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{u}{2\sigma^2}}$$

The moment generating function $\psi(s)$ is calculated as follows:

$$\begin{aligned} \psi(s) &= \int_{-\infty}^{\infty} e^{us} q(u) du \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_0^{\infty} e^{(s - \frac{1}{2\sigma^2})u} \frac{1}{\sqrt{u}} du \\ &= \frac{1}{\sqrt{1 - 2s\sigma^2}} \int_0^{\infty} \frac{1}{\sqrt{2\pi w}} e^{-\frac{w}{2}} dw \\ &= \frac{1}{\sqrt{1 - 2s\sigma^2}} \quad s < \frac{1}{2\sigma^2} \end{aligned}$$

In the third expression we make the substitution $\frac{w}{2} = (\frac{1}{2\sigma^2} - s)u$. The integral in the third line is recognized as integrating to 1, being, in fact, a special case of the density function $q(u)$ above. Notice that the integral and hence $\psi(s)$ exist only when $s < \frac{1}{2\sigma^2}$.

The log of the moment generating function and other useful functions can now be calculated. We have

$$\mu(s) = \log \hat{\nu}(s) = -\frac{1}{2} \log (1 - 2s\sigma^2)$$

$$\mu'(s) = \frac{\sigma^2}{1 - 2s\sigma^2}$$

$$\mu(s) - s\mu'(s) = -\frac{1}{2} \log (1 - 2s\sigma^2) - \frac{s\sigma^2}{1 - 2s\sigma^2}$$

$$\mu(s) - (s+1)\mu'(s) = -\frac{1}{2} \log (1 - 2s\sigma^2) - \frac{(s+1)\sigma^2}{1 - 2s\sigma^2}$$

$$\mu''(s) = \frac{\sigma^2}{(1 - 2s\sigma^2)^2} = 2(\mu')^2$$

Let $a_i =$ no of code points with i others within distance d . $p =$ expected no within distance d , on average

$$\text{Then } pm = \sum i a_i.$$

Now lets expurgate until $mp = 0$.

If we expurgate one with j others within distance d , we knock a_j out here. a_j is reduced by one. Also, the a_i for $i < j$ other points is reduced by one, for a reduction of another j . Total $2j$.

To reduce mp to zero would require no more than $\frac{mp}{2}$ expurgations, since each would reduce it by at least 2. Therefore expurgated code can have min distance D and still have

$$m - \frac{mp}{2} = m(1 - \frac{p}{2}) \text{ code points.}$$

$$p = e^{-R} \left(\frac{A}{P} \right)^{n/2} = m \left(\frac{A}{P} \right)^{n/2},$$

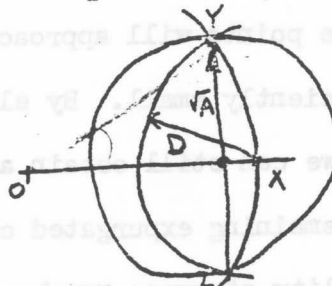
$$\text{So } m - \frac{mp}{2} = m - \frac{m^2}{2} \left(\frac{A}{P} \right)^{n/2}, \text{ which}$$

is maximum for $m \left(\frac{A}{P} \right)^{n/2} = p = 1$.

Then there are $\frac{m}{2}$ code points left after expurgation

Upper Bound on P_e for Gaussian Channel by Expurgated Random Code

In the gaussian channel with average power limitation we assume code words chosen at random in a sphere of radius \sqrt{P} . If the number of dimensions n is large enough, the fraction of points at a radius between $(1 - \delta)\sqrt{P}$ and \sqrt{P} will be greater than $1 - \epsilon$ for any positive ϵ and δ . We wish to calculate the rate $R = \frac{1}{n} \log M$ for a random code such that the expected number of code points within D of a given code point is less than or equal to one-half. In the figure



O is the origin, X is a code word at radius \sqrt{P} . The sphere of radius D centered on X intersects the original sphere of radius \sqrt{P} in an $(n - 1)$ sphere whose intersection with the plane of our drawing consists of the points Y and Z. All points interior to both spheres are included in the sphere of length OX and radius \sqrt{A} (in n dimensions). Hence, the volume common to the two spheres is less than or equal to the volume of this sphere, which is $K_n \sqrt{A}^n$ where K_n is the coefficient of r^n in the formula for the volume of an n -sphere. The total volume of the \sqrt{P} sphere is $K_n (\sqrt{P})^n$. If there are e^{nR} points chosen at random in the \sqrt{P} sphere, the expected number within distance D of one of the points, such as X, will be less than

$$e^{nR} \frac{K_n (\sqrt{A})^n}{K_n (\sqrt{P})^n}.$$



Now if $R = \log \sqrt{\frac{P}{A}} - \delta_1$, this expected number approaches zero as $n \rightarrow \infty$

for any $\delta_1 > 0$. If the point X is not on the surface of the \sqrt{P} sphere but at a slightly smaller radius, $\sqrt{P} - \epsilon$, the radius of the sphere, \sqrt{A} , is slightly larger, $\sqrt{A} + \delta_2$. However, by making ϵ approach zero, δ_2 approaches zero and its effect may be absorbed in δ_1 . Thus, if in our original sphere with points distributed at random we first eliminate all points except those within ϵ of the surface, the expected number within D of one of the points will approach zero as $n \rightarrow \infty$ provided the rate $R = \log \sqrt{\frac{P}{A}} - \delta_1$ and ϵ is sufficiently small. By eliminating those points which have neighbors within D we can still obtain a rate R as close as we wish to $\log \sqrt{\frac{P}{A}}$. Now, since in the remaining expurgated code no point has a neighbor closer than D , the probability of error may be calculated by our theorem on minimum separations. It will be less than e^{-nR} times the probability of noise carrying a point a distance $D/2$ or more. The distance $D/2$ can be related to

\sqrt{A} by the obvious trigonometric equation

$$\frac{D}{2\sqrt{P}} = \sin \frac{1}{2} \sin^{-1} \sqrt{\frac{A}{P}}$$

$$\Rightarrow A = D^2 \left(1 - \frac{D^2}{4P}\right)$$

Making use of the theorem on reliability for a given minimum separation, and the asymptotic formula for large n for $\text{erf } x$, we obtain

$$E \geq \frac{P}{2n} \sin^2 \frac{1}{2} \sin^{-1} \sqrt{\frac{A}{P}} - R.$$

Eliminating A by its relation to R , we get the final bound on reliability E

$$E \geq \frac{P}{2n} \sin^2 \frac{1}{2} \sin^{-1} e^{-R} - R.$$

Note that as $R \rightarrow 0$ this lower bound approaches $\frac{P}{4n}$, the same value as the upper bound on E previously derived. Thus we conclude that $E(0) = \frac{P}{4n}$.

Lower Bound on P_e in Gaussian Channel by Minimum Distance Argument

In a code of length n with M code words, let m_{is} ($i = 1, 2, \dots, M$, $s = 1, 2, \dots, n$) be the s^{th} coordinate of code word i . We are assuming an average power limitation P , so

$$\frac{1}{nM} \sum_{is} m_{is}^2 \leq P. \quad (1)$$

We also assume an independent Gaussian noise of power N added to each coordinate.

We now calculate the average squared distance between all the $M(M-1)/2$ pairs of points in n -space corresponding to the M code words. The squared distance from word i to word j is $\sum_s (m_{is} - m_{js})^2$. The average $\overline{D^2}$ between all pairs will then be

$$\overline{D^2} = \frac{1}{M(M-1)} \sum_{s,i,j} (m_{is} - m_{js})^2.$$

Note that each distance is counted twice in the sum and also that the extraneous terms included in the sum, where $i = j$, contribute zero to the sum. Squaring the terms in the sum,

$$\begin{aligned} \overline{D^2} &= \frac{1}{M(M-1)} \left[\sum_{ijs} m_{is}^2 - 2 \sum_s \sum_{ij} m_{is} m_{js} + \sum_{ijs} m_{js}^2 \right] \\ &= \frac{1}{M(M-1)} \left[2M \sum_{is} m_{is}^2 - 2 \sum_s \left(\sum_i m_{is} \right)^2 \right] \\ &\leq \frac{1}{M(M-1)} 2M P n M \\ \overline{D^2} &\leq \frac{2nMP}{M-1}, \end{aligned}$$

where we obtain the third line by using the inequality on the average power (1) and by noting that the second term is necessarily non-positive.

If the average squared distance between pairs of points $\leq 2nMP/M - 1$, there must exist a pair of points for whose distance this inequality holds. Each point in this pair is used $\frac{1}{M}$ of the time. The best detection for separating this pair (if no other points were present) would be by a plane normal to and bisecting the joining line segment, and either point would then give rise to a probability of error equal to that of the noise carrying a point half this distance or more in a specified direction. We arrive therefore, at a probability of error

$$P_e \geq \frac{1}{M} \Pr \left[\text{noise in a certain direction} > \frac{1}{2} \sqrt{\frac{2nMP}{M-1}} \right]$$

$$P_e \geq \frac{1}{M} \operatorname{erf} \left(\sqrt{\frac{Mn^2}{(M-1)2N}} \right).$$

As $n \rightarrow \infty$ and assuming $M \rightarrow \infty$ also in such a way as to approach a definite rate $\frac{1}{n} \log M \rightarrow R > 0$ we may translate this into a bound on the asymptotic reliability. This is done by using the asymptotic formula, $\operatorname{erf} x \rightarrow \frac{1}{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. Using this, taking the logarithm and dividing by n gives the simple upper bound on reliability

$$E \leq \frac{P}{4N}.$$

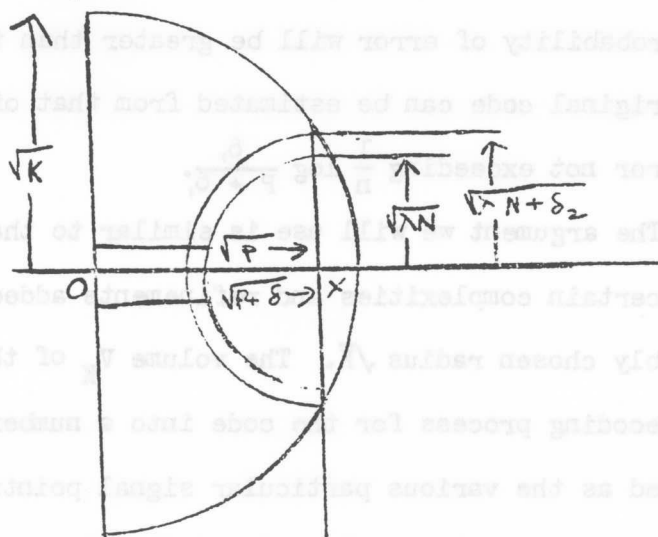
The Sphere Packing Bound for the Gaussian Power Limited Channel

The analog of the sphere packing argument can be carried out in an interesting geometrical fashion for the gaussian channel. We assume an average power limitation P and an independent gaussian noise of n coordinates with variance N in each coordinate. Consider the n -sphere S whose squared radius is $P + \delta_1$. Since the average squared radius to the signal words is P or less, a fraction at least $\frac{\delta_1}{P + \delta_1}$ of these words are within the $P + \delta_1$ sphere for, if not, the fraction greater than $(1 - \frac{\delta_1}{P + \delta_1})$ at distance at least $P + \delta$ would give more than P for the contribution to the average power by themselves. We will estimate the errors due to only the signal words inside the $P + \delta_1$ sphere. Even if all code words outside this sphere never caused errors and this minimum possible fraction $\frac{\delta_1}{P + \delta_1}$ were inside the sphere, the probability of error for the entire code would be that of the code consisting of these interior points multiplied by $\frac{\delta_1}{P + \delta_1}$, and in general the probability of error will be greater than this. Thus the reliability of the original code can be estimated from that of the interior points with an error not exceeding $\frac{1}{n} \log \frac{\delta_1}{P + \delta_1}$.

The argument we will use is similar to that in the discrete channel but with certain complexities and refinements added. We consider a sphere of suitably chosen radius \sqrt{K} . The volume V_K of this sphere will be divided by the decoding process for the code into a number of regions, regions which are decoded as the various particular signal points. To each signal point we will assign a certain volume V_1 of "high" probability density and a second volume V_2 of "low" probability density. These regions V_1 and V_2 are congruent for the different signal points. The probability density of a point being

carried by noise into any part of its V_1 region will be greater than the density for any part of its V_2 region. Both of these regions will, for any signal point, lie entirely within the sphere of radius \sqrt{K} . The conclusion will be that for any placing of V_K/V_1 points the probability of error will be at least equal to the probability of a point being carried into V_2 region. This is because, in a way similar to the discrete process, starting with the original partitioning of V_K , we can reallocate volume assigned to a given point in order of decreasing probability density and equalize allocation between points until each point has V_1 assigned to it. These operations preserve total volume and decrease (calculated) probability of error. When the equalization is complete, each signal point has its V_2 region assigned entirely to other points, and consequently the probability of error is at least that of a point being taken to its V_2 region.

In the figure



O is the origin, X is a signal point at maximal radius $\sqrt{P + \delta_1}$, and the large circle is the intersection of the K sphere with the plane of the drawing. At X we construct the hyperplane perpendicular to OX, and let the distance from

X to the intersection of this plane with the K sphere be $\sqrt{\lambda N + \delta_2}$. Here, N is the average noise power, λ is an arbitrary multiplier, and δ_2 is a small quantity which will eventually approach zero. Now construct the two hemispheres of radii $\sqrt{\lambda N}$ and $\sqrt{\lambda N + \delta_2}$ centered on X, pointed toward O and bounded by the hyperplane. It is clear that the entire volumes of both of these hemispheres are within the large K sphere. The smaller hemisphere is the V_1 region for signal point X and the shell between the hemispherical surfaces is the V_2 region. For any other signal point, a similar pair of hemispheres is constructed by drawing the line from the origin to the signal point, constructing the perpendicular hyperplane and constructing hemispheres of radii $\sqrt{\lambda N}$ and $\sqrt{\lambda N + \delta_2}$, facing toward the origin. If the origin itself were a signal point, any hyperplane through the origin may be used. It is obvious in the drawing that any point of these hemispheres actually in the plane of the drawing is within the K sphere (being nearer to the origin than \sqrt{K}). But the plane of the drawing may be made to pass through any desired point in the hemisphere by suitable rotation, hence the property is true in general.

Since probability density for a given displacement from a signal point is a monotone decreasing function of the actual distance of displacement, the probability density for any point in the shell is less than that for any point in the inner hemisphere. Let M_0 be the number of signal points such that the combined volume of their small hemispheres is just equal to that of the K sphere. Thus

$$M_0 = \frac{2\sqrt{K}^n}{\sqrt{\lambda N}^n} = 2 \left(\frac{\sqrt{K} + \delta_1 + \delta_2}{\sqrt{\lambda N}} \right)^n$$

Now, whatever the decoding system or the placement of M_0 points interior to the $\sqrt{P + \delta_1}$ sphere, the probability of error P_e (due only to errors inside the K sphere) will exceed the probability of a point being carried into its V_2 shell. This follows from our general argument concerning reallocation of volume in accordance with higher probability. Thus if the message placement and decoding system allocate any volume in shells or other low probability density regions to code points, a lower calculated P_e would occur if this were calculated as though at the higher probability density of the inner hemisphere. When this reallocation is finished, we have a probability of error satisfying

$$P_e \geq \frac{1}{2} \Pr \left[\lambda N < Z < \lambda N + \delta_2 \right]$$

where Z is the squared radial displacement of a point due to noise (divided by n). Since Z is the sum of n independent Gaussian variates each with variance N , Z is distributed (apart from scale) according to the χ^2 distribution with n degrees of freedom. Thus

$$P_e \geq \frac{1}{2} \int_{\lambda}^{\lambda + \frac{\delta_2}{N}} f_n(\chi^2) d\chi^2$$

$$= \frac{1}{2} \int_{\lambda}^{\lambda + \frac{\delta_2}{N}} \frac{\left(\frac{\chi^2}{2}\right)^{\frac{n}{2}-1}}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} e^{-\frac{\chi^2}{2}} d\chi^2.$$

For any given $\delta_2 > 0$, the logarithm of the χ^2 distribution from λ to $\lambda + \frac{\delta_2}{N}$ is asymptotic to $\frac{n}{2} \log \frac{e^{\lambda-1}}{\lambda}$. This can be easily shown by use of the moment generating function and the results on the tails of distributions obtained previously. Consequently our reliability E , as $n \rightarrow \infty$, is asymptotically less than or equal to $\frac{1}{2} \log \frac{e^{\lambda-1}}{\lambda}$. Also the rate $R = \frac{1}{n} \log M \frac{1}{n} \log \frac{P + \delta_1}{\delta_1} \rightarrow \frac{1}{2} \log$

$\frac{P + \lambda N + \delta_1 + \delta_2}{\lambda N}$. Since this is true for any $\delta_1, \delta_2 > 0$, we may omit them

entirely and obtain asymptotic bounds for E and R as follows.

$$E \leq \frac{1}{2} \log \frac{e^{\lambda-1}}{\lambda}$$

$$E \geq \frac{1}{2} \log \frac{4e^{\frac{\lambda}{4}-1}}{\lambda}$$

$$R \leq \frac{1}{2} \log \left(1 + \frac{P}{\lambda N} \right).$$

These formulas give an upper bound on the reliability curve in a parametric form using the parameter λ which ranges from 1 to ∞ . With λ just greater than 1, we have a rate just below channel capacity and a reliability bound which is just slightly positive. As the value of λ increases, the rate R decreases and the bound on E increases, becoming infinite when λ is infinite and the bound on rate is zero. Of course the bound based on minimum distance shows that the actual E curve does not exceed $\frac{P}{LN}$ as $R \rightarrow 0$.

The T-terminal Channel

Almost all previous work on coding theory has dealt with a one-directional channel having an input or transmitting point and an output or receiving point, or, at most, with this arrangement plus a feedback channel from the receiving point to the transmitting point whose function was thought of as a possible aid in forward communication. Many cases arise, however, in which a number of information terminals are involved and both backward and forward communication is of interest perhaps between all pairs of terminals. As examples we may cite telephony (or even ordinary direct conversation) where communication in both directions is important, or a network of radio or television stations in which there are a number of communication links using a common medium.

A further complication is introduced by the possibility of competition or conflicting interest among the individuals controlling the operation of the various terminals. As an example we have the case of a secrecy system which is best thought of as a three-terminal channel with the transmitter as one input, a receiver as one output and the enemy cryptanalyst as a second output. The object is to transmit information from the transmitter to the receiver without knowledge by the enemy. A second example is the problem of "jamming", again a three-terminal channel, but now the enemy has an input rather than an output and his object is to reduce or eliminate the direct transmission of information.

These possibilities suggest that we should frame general definitions of T-terminal channels and study their characteristics from the information theoretic point of view. We shall here, for simplicity, limit ourselves to the discrete case quantized in time.

Definition: A T-terminal finite state channel consists of T inputs x_i ($i = 1, 2, \dots, T$) each of which may assume values from a finite alphabet (not necessarily the same for the different inputs), T outputs y_1, y_2, \dots, y_T each of which can assume values from an associated finite alphabet, and a state variable S which can assume any of a finite set of values $1, 2, \dots, D$. Finally, there are conditional probabilities for the next outputs and the next state conditional on the current inputs and current state:

$$Pr(y_i/S, x_1, x_2, \dots, x_T) \text{ and } Pr(S'/S, x_1, x_2, \dots, x_T)$$

Definition: A memoryless T-terminal finite state channel is one in which the state S can assume only a single value.

Definition: A noiseless T-terminal discrete channel is one in which all probabilities are either 0 or 1. Thus, the next state and the next outputs are strictly determined by the current state and current inputs. In the noiseless memoryless case, this state can have only one value so the next outputs are functions of the current inputs.

In operation of a T-terminal channel we imagine operators or equipment at each of the terminals. Also at each terminal, in general, will be an information source. The operators are attempting to transmit information produced by the sources between the terminals according to some general plan and system of codes which has been agreed upon. In general, the operator at terminal i can control the input i but only as a function of the data available to him at the time. This includes the past and present of output i and the output of message source i up to the present time but not the future of these random functions, nor any of the other inputs, outputs or message sources (past or future).

We will first consider the completely cooperative situation in which the operation of all terminals is directed toward a common end. The problem is very similar to a one-person game in the game theoretic sense with "split personality" for the player. We can think of the operators at the various terminals conferring at the beginning on a general strategy, selection of codes and decoding operations, and then going to their respective terminals and operating the system according to the agreed-upon plan. Together they act like a single player whose knowledge in making different moves is not coextensive.

In the more general case, one may consider a p -person game in which the T -terminals are partitioned into p subsets, the operators in each subset having a common purpose which may conflict with those of other subsets. The operators in a given subset agree on a strategy to promote their goals and act as one person in a kind of p -person game.

In the fully cooperative case there are many utilities one might wish to maximize in a given channel. In line with basic coding theory, however, our attention is directed to the question of generalizing the coding theorem for a noisy channel to this kind of a situation. In other words, we would like to find the capabilities and limitations of a T -terminal channel with regard to essentially errorless transmission of information between the different terminals. At a given terminal, say terminal 1, we may imagine that the information source 1 produces information which is destined for various other terminals 2, 3, ..., T . It might also produce some information which was intended for both terminals 2 and 3, and some for both 2 and 4, etc., and indeed it might have a component intended for any subset of the other terminals. The same may of course be said of any other terminal. In general,

we think of each message source as producing not one but 2^{T-1} streams of independent information intended for the 2^{T-1} subsets (omitting the null subset) of the other $T-1$ terminals.

A simple two-terminal one-way channel is characterized at the simplest coding level by its capacity C . In the T -terminal case, we must consider the capacities of all the different types just described, that is, C_{iK} , the capacity from terminal i to subset K of the remaining terminals, a total of $T(2^{T-1} - 1)$ different capacities. Furthermore, these are not fixed quantities but, in general, capable of some variability. Thus, one may increase one of these capacities at the expense of reducing another. Our fundamental problem is not to evaluate a single C as before but to find which sets of values of C_{iK} are possible.

In the case of only two terminals but with an input and output at each terminal, there are only two different capacities C_{iK} , since there is only one non-null subset of the remaining terminals. These capacities we may write C_{12} and C_{21} . Our problem is to find the possible values of the pair (C_{12}, C_{21}) or, better, the boundary of this domain in the C_{12}, C_{21} space. This boundary may be called the capacity surface.

The channel in Fig. 1 is a simple example where the two boxes represent an ordinary one-way memoryless channel with capacities C_1 and C_2 . The graph at the right of Fig. 1 shows the region of attainable rates in the two directions and the heavy line boundary of this is the capacity surface. In this case transmission in either direction neither aids nor hinders transmission in the reverse direction (feedback cannot increase forward transmission in a memoryless channel).

The channel in Fig. 2 is more interesting from this point of view. The two binary inputs from the two terminals are added mod 2 and the output is a common output going to both terminals. Here again it is possible to achieve points in a rectangle. Note that at each transmitter the transmitter symbol should be added mod 2 to the next received symbol to compensate for its effect. It is curious that, in a sense, two bits per time interval are going through the vertical line of the drawing, one destined in each direction.

Another channel is indicated in Fig. 3. There are three input letters a, b, c at the left terminal and three input letters A, B, C at the right terminal. If a is used at the left, the channel from the right is as shown in the figure, a channel with capacity 1. B or C come through to corresponding received letters B' and C' while A divides with probability $\frac{1}{2}$ between these. If b or c is used, the channel from right to left has zero capacity, all letters A, B, C dividing equally between B' and C'. In the reverse direction, the situation is similar with capital letters exchanged for small letters. Thus there is a direct conflict between sending information to the right or the left. Any point in the triangular region can be attained but, we suspect, nothing outside. To obtain a point on the diagonal boundary, say $C_{12} = x$ and $C_{21} = 1 - x$, the channel may be used x of the time to the right (that is, the right hand operator uses A) and $1 - x$ of the time to the left (the left hand operator uses a). In each case, the other operator sends at full capacity.

In the general T-terminal memoryless channel, essentially this apportionment of time may be carried out to prove the following theorem.

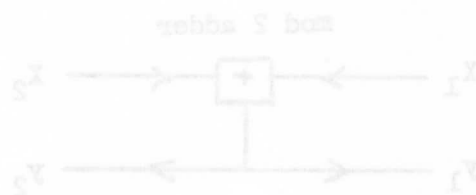
Theorem: The capacity surface is convex outward. That is, if the sets $C_{i\sigma}$ and $C'_{i\sigma}$ can be attained (where i ranges over the terminals and σ over subsets of terminals excluding i), then the set of capacities

$$C''_{i\sigma} = \lambda C_{i\sigma} + (1 - \lambda) C'_{i\sigma}$$

$$0 \leq \lambda \leq 1$$

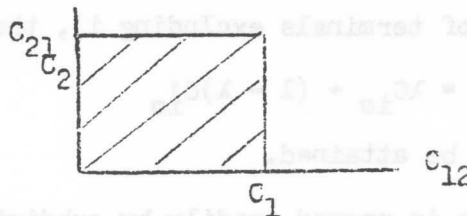
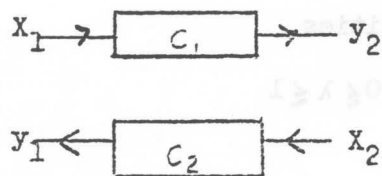
can also be attained.

This is proved readily by subdividing the time between the coding systems which give $C_{i\sigma}$ and $C'_{i\sigma}$ in the ratios λ and $1 - \lambda$. If these are irrational, they may of course be approximated by a sequence of rationals.



Theorem: The capacity surface is convex outward. That is, if the sets

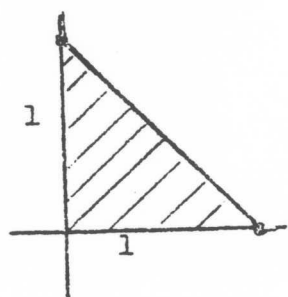
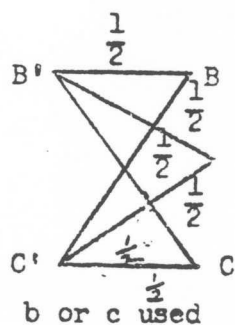
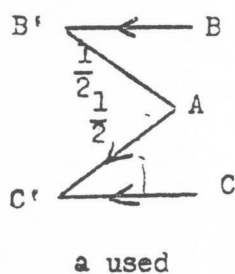
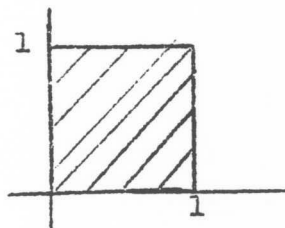
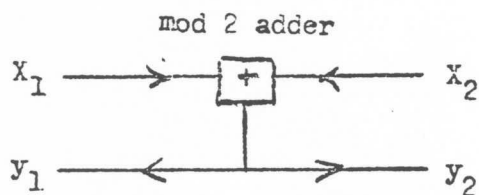
C_{12} and C_{21} can be attained (where λ ranges over the terminals and λ over



This is proved readily by subdividing the line between the coding sys-

tems which give C_{12} and C_1 in the ratios λ and $1-\lambda$. If these are irrational,

they may of course be approximated by a sequence of rationals.



1:66

Conditions for Constant Mutual Information

Theorem: In a channel with $p_i(j)$ matrix and P_i input probabilities necessary and sufficient conditions that the mutual information be constant are that

$$(1) \quad p_i(j) = f_j, \text{ a function of } j \text{ only}$$

$$(2) \quad \sum_{S_j} P_i = h, \text{ independent of } j, \text{ when } S_j \text{ is the set of input letters that can cause output letter } j.$$

We also have $\sum_j f_j = h^{-1} = e^{-I}$, where I is the constant information value.

Proof: Suppose $\log \frac{P_i p_i(j)}{P_i Q_j} = I$. Then $p_i(j) = e^I Q_j$ a function of

only. Also if $q_j(i)$ is the conditional probability of i given j , then

$$\frac{Q_j q_j(i)}{P_i Q_j} = e^I$$

$$q_j(i) = e^I P_i$$

$$1 = \sum_{S_j} q_j(i) = e^I \sum_{S_j} P_i$$

To prove the sufficiency, assume (1) and (2). From (1)

$$\frac{P_i p_i(j)}{P_i Q_j} = \frac{f_j}{Q_j} = \lambda_j = \frac{Q_j q_j(i)}{P_i Q_j} = \frac{q_j(i)}{P_i}$$

Now summing $P_i \lambda_j = q_j(i)$ over $i \in S_j$ and using (2),

$$h \lambda_j = 1$$

so λ_j is h^{-1} independent of j . Hence $I = \log h^{-1}$.

Simple Proof that $H_{Y|X}(x) \leq H(x)$

We wish to prove that

$$\sum_{i,j} p(i,j) \log p_i(j) \leq - \sum_j p(j) \log p(j)$$

We will prove this for each particular j ; summing on j will then give the desired result. Thus we will show

$$- \sum_i p(i,j) \log p_i(j) \leq - p(j) \log p(j)$$

or

$$- \sum_i p(i) p_i(j) \log p_i(j) \leq - \sum_i p(i) p_i(j) \log \sum_i p(i) p_i(j)$$

Consider $\phi(x) = x \log x$. This function is convex downward for $x \geq 0$ since

$\phi''(x) = \frac{1}{x} > 0$. Therefore it satisfies the inequality (see Hardy, Littlewood

and Polga "Inequalities" p. 74)

$$\phi\left(\sum_i q_i x_i\right) \leq \sum_i q_i \phi(x_i) \quad \text{where } \sum_i q_i = 1$$

Take $x_i = p_i(j)$ and $q_i = p(i)$

$$\sum_i p(i) p_i(j) \log \sum_i p(i) p_i(j) \leq \sum_i p(i) p_i(j) \log p_i(j)$$

This is, after multiplication by (-1) and summation on j , the desired inequality.

Equality occurs only if all $p_i(j)$ for a given j are equal. Then $p_i(j) = q(j)$

and $p(i,j) = p(i) q(j)$. That is, the two events are independent.

The Central Limit Theorem with Large Deviations

The central limit theorem states that under certain general conditions the sum of n independent random variables is approximately gaussian in the neighborhood of its mean value when n is large. The most common theorems of this class give good estimates of the probability at deviations of the order of $K\sqrt{n}$ from the mean, while more advanced results with added terms (for example, the results on p. 147 of Feller, Probability Theory and Its Applications) allow somewhat larger deviations but still require that the deviation from the mean divided by n approach zero for the estimate to be asymptotic to the correct value with large n .

We will develop asymptotic formulas under certain conditions for the probability density, the probabilities of the tails of the distributions, etc., for arbitrary deviations. In the usual central limit theorem, the behavior near the mean is related to the characteristic functions or, as we prefer here, the moment-generating functions near the value zero. It is interesting that the results here show that the distribution remote from the mean is in a very similar fashion related to the moment-generating functions at arguments away from zero. Thus we are able to attach a fairly direct significance to the value and derivatives of the moment-generating functions at non-zero arguments. Indeed, the method of derivation of our asymptotic estimates is a kind of manipulation trick whereby points away from zero are translated into zero. This device is due to Escher and has been used by Cramér in a manner similar to our analysis. However our results go further than those of Cramér, most of whose work applied only near the mean of the distribution.

Let $F(x) = \Pr \{u \leq x\}$ be the distribution function for the random variable u . The moment-generating function is then

$$\phi(s) = \int_{-\infty}^{\infty} e^{sx} dF(x)$$

Let this converge in the range $A < s < B$ (either or both A and B may be infinite).

We are interested only in cases where $B > 0$. This includes distribution functions which are bounded in range or which approach zero and one exponentially or faster, as with the gaussian distribution or the distribution whose density is $e^{-\frac{1}{2}|x|}$.

The moment-generating function is an analytic function of s (thought of as a complex variable) in the strip where $A < \text{Re}[s] < B$. If n variables, all independent and distributed according to the same $F(x)$, are added, the sum X is distributed according to the n -fold convolution $F_n(x)$. The moment-generating function of $F_n(x)$ is

$$\Phi(s) = [\phi(s)]^n.$$

We wish to estimate $F_n(\lambda n)$ when n is large.

Consider a new random variable u whose distribution function $G(z)$ is defined by

$$G(z) = \frac{\int_{-\infty}^z e^{s(z)} dF(z)}{\int_{-\infty}^{\infty} e^{s(z)} dF(z)}$$

i.e.,

$$dG(z) = \frac{e^{s(z)} dF(z)}{\int_{-\infty}^{\infty} e^{s(z)} dF(z)}$$

Here s_0 is an arbitrary real constant lying between B and A .

The moment-generating function for $G(\bar{x})$ is

$$\psi(s) = \int_{-\infty}^{\infty} e^{s\bar{x}} dG(\bar{x})$$

$$\frac{\int_{-\infty}^{\infty} e^{s\bar{x}} dF(\bar{x})}{\int_{-\infty}^{\infty} dF(\bar{x})}$$

$$\frac{\psi(s+s_0)}{\psi(s_0)}$$

The mean and variance of the G distribution may be found from the first and

second derivatives of $\psi(s)$ evaluated at $s = 0$. Thus

$$\bar{x} = \psi'(0) = \frac{\psi'(s_0)}{\psi(s_0)}$$

$$\sigma^2 = \psi''(0) - [\psi'(0)]^2 = \frac{\psi''(s_0)}{\psi(s_0)} - \left[\frac{\psi'(s_0)}{\psi(s_0)} \right]^2$$

Now suppose n variables, all independent and distributed according to $G(\bar{x})$ are added. The sum z will be distributed according to $G_n(\bar{x})$ with the moment-generating function

$$\bar{\psi}(s) = [\psi(s)]^n = \left[\frac{\psi(s+s_0)}{\psi(s_0)} \right]^n$$

This implies that

$$dG_n(z) = \frac{e^{s_0 z}}{[\psi(s_0)]^n} dF_n(z),$$

since an addition of s_0 in the argument of the generating function corresponds to a multiplication by $e^{s_0 z}$ in the distribution function.

Thus the distribution $G(z)$ after n -fold convolution is still closely related to the n -fold convolution of $F(x)$.

$$dF_n(x) = \phi(s_0)^n e^{-s_0 x} dG_n(x)$$

The basic method of using this relation to study the behavior of the distribution $F_n(x)$ is as follows. A value of s_0 is chosen in such a way as to make the mean of the G distribution occur at the value x of F_n in which we are interested. When this is done, $G_n(x)$ can be estimated well from the ordinary central limit theorems, since these are particularly good at and near the mean. The relation between F_n and G_n is then used to translate estimates of G_n behavior into estimates of F_n behavior.

It is convenient to use in place of the moment-generating function $\phi(s)$ its logarithm, which we will denote by $\mu(s)$. This function is sometimes called the semi-invariant generating function. In terms of $\mu(s)$ we have

$$dF_n(x) = e^{n\mu(s)} e^{-sx} dG_n(x).$$

The successive derivatives of $\mu(s)$ evaluated at zero are called the semi-invariants of the F distribution. In particular,

$$\mu(0) = 1$$

$$\mu'(0) = \int x dF(x) = \text{mean of } F \text{ distribution}$$

$$\mu''(0) = \int x^2 dF(x) - \mu'(0)^2 = \sigma^2 \text{ of } F \text{ distribution}$$

For the $G(x)$ distribution, the log moment generating function $\mu_G(s)$ is given by (taking the logarithm of (1))

$$\mu_G(s) = \mu(s + s_0) - \mu(s_0).$$

Consequently, for all derivatives (using a superscript to denote differentiation)

$$\mu_G^{(n)}(0) = \mu^{(n)}(s_0).$$

In words, the semi-invariance of the G distribution are the derivatives of the F distribution evaluated at s_0 . In particular, the mean and variance of the G distribution are $\mu'(s_0)$ and $\mu''(s_0)$. The mean and variance of the G_n distribution are, similarly, $n\mu'(s_0)$ and $n\mu''(s_0)$.

Note that the operation of forming the new distribution function $G(x)$ (or the corresponding new random variable) from a given distribution function $F(x)$ (or its random variable) is a group operation. Thus, if we let T_s denote the operation which applied to $F(x)$ gives $G(x)$,

$$T_s F(x) = G(x) = \int_{-\infty}^x e^{+sx} dF(x) / \int_{-\infty}^{\infty} e^{+sx} dF(x),$$

then the T_s form an additive Abelian group isomorphic to the additive group for real numbers,

$$T_{s1} \cdot T_{s2} = T_{s1 + s2}$$

$$T_0 = I.$$

The operation T_s is distributive over the binary operation of convolution (which itself is commutative and associative). Thus, if we denote convolution of two distribution functions by an asterisk and repeated convolution of the same distribution by an asterisk preceding the exponent, we have

$$T_s (F * G) = (T_s F) * (T_s G)$$

$$T_s (F^{*n}) = (T_s F)^{*n}.$$

This last equation, when we operate on both sides by T_{-s} , gives the basic result we have used in estimating tails of distributions,

$$F^{*n} = T_{-s} (T_s F)^{*n}.$$

If we think of the operation $T_S^F = G$ as producing a new probability measure for the random variable x , then there is a one to one correspondence between points in the two probability spaces involved, the F space and the G space, and also between points in the product spaces of F with itself n times and G with itself n times. The probability measures in the two spaces are very closely related. If a point in the F space has value x and probability P , the corresponding point in the G space has value x and probability $Q = e^{sx} P / \int e^{sx} dF(x)$. If we select a subset S_1 of points whose x values all lie between A and B , then we will have

$$k e^{sB} \text{Prob}_F[S_1] \leq \text{Prob}_G[S_1] \leq k e^{sA} \text{Prob}_F[S_1]$$

$$\text{where } k^{-1} = \int e^{sx} dF(x).$$

The Chernoff Inequality

To illustrate the use of the G distribution in estimating the tail of the F_n distribution, we will first give a crude but simple and useful bound on the tail due to Chernoff, who proved it by a different method. We have

$$F_n(x) = e^{n\mu(s_0)} \int_{-\infty}^x e^{-ys_0} dG_n(y).$$

If $s_0 < 0$, the maximum of e^{-ys_0} occurs at $y = x$. Thus

$$F_n(x) \leq e^{n\mu(s_0)} e^{-xs_0} \int_{-\infty}^x dG_n(x) \quad s_0 < 0$$

$$\leq e^{n\mu(s_0)} e^{-xs_0}$$

This is true for any x and any s_0 , but to obtain the most favorable bound we should choose s_0 so as to minimize $n\mu(s_0) - xs_0$ (for the x in question). Remembering that $\mu(s)$ is analytic and that $\mu''(s) > 0$ (since it is a variance) the necessary and sufficient condition for a minimum is that $n\mu'(s_0) = x$. This will have a unique solution in s_0 . However, it is more convenient to express our result parametrically in terms of s , or, dropping the subscript, in terms of s . Thus

$$F_n(n\mu'(s)) \leq e^{n(\mu(s) - s\mu'(s))} \quad s < 0$$

In a similar fashion, by integrating from x to ∞ , we obtain a bound on the tail in the positive direction of exactly the same type. Combining these results we have the following. If $F_n(x)$ is the distribution function of the sum of n identically distributed random variables, each with log moment generating function $\mu(s)$ which exists for $A < s < B$, then

$$F_n(n\mu'(s)) \leq e^{n(\mu(s) - s\mu'(s))} \quad A < s < 0$$

$$1 - F_n(\eta_n(s)) \leq e^{n(\mu(s) - s\mu'(s))} \quad 0 < s < B$$

These bounds are very useful in that they are extremely simple to compute. Furthermore, while they are not asymptotic to F_n or $1 - F_n$ as $n \rightarrow \infty$, the logarithms of the bounds are asymptotic to the logarithms of F_n and $1 - F_n$ (in the respective s ranges, as will be seen later). Hence if we are interested only in the logarithm of F_n for large n , the Chernoff bounds give the correct asymptotic behavior.

In the succeeding sections we will make a more detailed estimation of $F_n(x)$ and $1 - F_n(x)$ by using more care in estimating the integral above. This will lead to sharper upper bounds, to lower bounds, and to asymptotic values of F_n distributions. Also, we will see how to obtain the most favorable bounds. This is true for any x and any n , but to obtain the most favorable bounds is sometimes easier than it is.

we should choose s_0 so as to minimize $\eta(s_0) - xs_0$ (for the x in question). Remembering that $\eta(s)$ is analytic and that $\eta'(s) > 0$ (since it is a variance), the necessary and sufficient condition for a minimum is that $\eta'(s_0) = x$.

This will have a unique solution in s_0 . However, it is more convenient to express our result parametrically in terms of s , or, dropping the subscript,

$$F_n(\eta(s)) \leq e^{n(\mu(s) - s\mu'(s))} \quad \text{in terms of } s. \quad \text{Thus}$$

In a similar fashion, by integrating from x to ∞ , we obtain a bound on the tail in the positive direction of exactly the same type. Combining

these results we have the following. If $F_n(x)$ is the distribution function of the sum of n identically distributed random variables, each with log

$$F_n(\eta(s)) \leq e^{n(\mu(s) - s\mu'(s))} \quad \text{moment generating function } \eta(s) \text{ which exists for } 1 \leq s < B, \text{ then}$$

Upper and Lower Bounds on the Tails of Distributions

Theorem: The distribution of the sum of n identically distributed independent random variables satisfies

$$\left\{ \begin{array}{l} F_n(n\mu'(s)) \\ 1 - F_n(n\mu'(s)) \end{array} \right\} \leq \frac{1}{|s| \sqrt{2\pi n\mu''(s)}} e^{n(\mu(s) - s\mu'(s))} \left(1 + \frac{1}{s^2 n\mu''(s)} + 2c \sqrt{\mu''(s)} \left(\frac{\mu'''(s)}{\mu''(s)} + \frac{3\mu''''(s)}{\mu''(s)^2} \right) \right)$$

where $\mu(s)$ is the log moment generating function of $F(x)$; μ' , μ'' , and μ''' are derivatives and c is an absolute constant, the constant in the Berry theorem relating to the approximation in the central limit theorem with error less than or equal to $\frac{c\beta_3}{\sigma^3\sqrt{n}}$. Also c may be replaced in the inequality by $3 \ln n / \sqrt{n}$.

Proof: We have

$$1 - F_n(n\mu'(s)) = e^{n\mu(s)} \int_{n\mu'(s)}^{\infty} e^{-sx} dG_n(x) \quad s > 0$$

On making the substitution

$$y = \frac{x - n\mu'(s)}{\sqrt{n\mu''(s)}}$$

and writing $H_n(y)$ for $G_n(\sqrt{n\mu''(s)} y + n\mu'(s))$, we obtain an H_n distribution, with mean at zero and variance one, suitable for application of ordinary central limit results. The equality above becomes

$$F_n(n\mu'(s)) = e^{n(\mu(s) - \mu'(s)s)} \int_0^{\infty} e^{-s\sqrt{n\mu''(s)} y} dH_n(y).$$

$H_n(y)$ can be estimated from the Cramér-Berry-Esseen theorem. Thus

$$H_n(y) = \Phi(y) + B(y)$$

$$B(y) < \frac{c \rho_3}{\sqrt{n}}$$

where $\rho_3 = \beta_3 / \mu''^{3/2}$ and β_3 is the third absolute moment of F .

The integral then breaks into two parts. First we have:

$$\begin{aligned}
 \int_0^{\infty} e^{-s \sqrt{n\mu''}(s)} y d\Phi(y) &= \frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{-y^2/2 - s \sqrt{n\mu''}(s)} y dy \\
 &= \frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{-\frac{(y + s \sqrt{n\mu''}(s))^2}{2}} e^{\frac{s^2 n\mu''(s)}{2}} dy \\
 &= e^{\frac{s^2 n\mu''(s)}{2}} \Phi(s \sqrt{n\mu''}(s)) \\
 &\leq e^{\frac{s^2 n\mu''(s)}{2}} \frac{e^{-\frac{s^2 n\mu''(s)}{2}}}{\sqrt{2\pi} s \sqrt{n\mu''}(s)} \left(1 + \frac{1}{s^2 n\mu''(s)}\right) \\
 &= \frac{1}{s \sqrt{2\pi n\mu''}(s)} \left(1 + \frac{1}{s^2 n\mu''(s)}\right)
 \end{aligned}$$

The second integral involving $\delta B(y)$ may be bounded by integrating by parts.

$$\begin{aligned}
 \int_0^{\infty} e^{-s \sqrt{n\mu''}(s)} \delta B(y) &= e^{-s \sqrt{n\mu''}(s)} y B(y) \Big|_0^{\infty} + s \sqrt{n\mu''}(s) \int_0^{\infty} B(y) e^{-s \sqrt{n\mu''}(s)} dy \\
 &\leq \frac{cp_3}{\sqrt{n}} + s \sqrt{n\mu''}(s) \frac{cp_3}{\sqrt{n}} \frac{1}{s \sqrt{n\mu''}(s)} \\
 &= \frac{2cp_3}{\sqrt{n}} \\
 &= \frac{2c}{s \sqrt{2\pi n\mu''}(s)} \frac{\beta_3}{\mu''^{3/2}}
 \end{aligned}$$

Collecting these terms, we obtain a bound for the tail of the distribution:

$$F_n(n\mu') \leq \frac{1}{s \sqrt{2\pi n\mu''}} e^{n(\mu - s\mu')} \left[1 + \frac{1}{s^2 n\mu''} + 2c \frac{\beta_3}{\mu''^{3/2}} s \sqrt{2\pi n\mu''} \right]$$

By a well-known inequality $\beta_3^{1/3} < \beta_4^{1/4}$ and $\beta_4 = \mu^{1/4} = 3(\mu')^2$.
 Consequently $\rho_3 \leq \frac{\beta_4^{3/4}}{(\mu')^{3/2}} = \left[\frac{\mu^{3/4} + 3(\mu')^2}{(\mu')^2} \right]^{3/4}$. This results in the
 final inequality involving only n , s and μ and its derivatives (together
 with the unknown absolute constant c). Since the original Lyapunov's
 theorem (with constant estimated by Cramer) gives an inequality for $B(y)$
 as follows

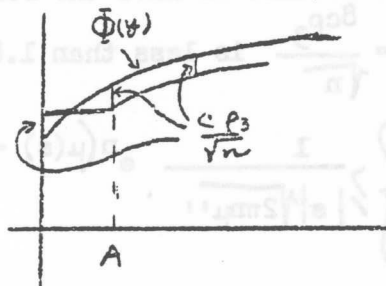
$$B(y) < \rho_3 \left(\frac{3 \log n}{n} \right)$$

we may, in our inequalities, replace c by $3 \log n$. This makes them com-
 pletely definite, although at a certain loss in order of magnitude as a
 function of n when n is large.

To estimate a lower bound on the tail, the method is identical up to
 the point where we must estimate the following integral,

$$\int_0^{\infty} e^{-s \sqrt{n \mu'}} y \, dH_n(y).$$

Again using the theorem involving $\frac{c\rho_3}{n}$, it is evident that the monotone
 increasing function $H_n(y)$ which would minimize this integral, subject to
 being within $\frac{c\rho_3}{n}$ from $\Phi(y)$, would be that shown in the figure.



This function starts at zero as high above $\Phi(y)$ as possible and is con-
 stant at this value as long as possible. It then increases as slowly as
 possible. It is easily shown that any other permissible $H_n(y)$ gives a

larger integral than this function; when changed into this function in the obvious way the integral is decreased. In the figure the corner in the curve occurs at A, which is such that $\Phi(A) - \Phi(0) = \frac{2c\rho_3}{\sqrt{n}}$. To obtain a simple estimate of A which is on the safe side (that is, larger than the actual A) we may approximate $\Phi(y)$ by a straight line passing through $\frac{1}{2}$ at $y = 0$ and of slope $\frac{1}{4}$. This will lie below $\Phi(y)$ out to $y = 1.86$. Hence if the A computed from this straight line, namely $A = \frac{8c\rho_3}{\sqrt{n}}$, is less than or equal to 1.86, the estimate is safe. If not, the more elaborate formula involving $\Phi(A)$ may be used. In any case, our lower bound integral becomes

$$\int_A^{\infty} e^{-s\sqrt{n\mu''}} y d\Phi y.$$

On completing the square, as before, this becomes

$$\begin{aligned} \frac{s^2 n\mu''}{e^2} \Phi(s\sqrt{n\mu''} - A) &\geq \frac{s^2 n\mu''}{e^2} - \frac{(s\sqrt{n\mu''} - A)^2}{2} \frac{1}{(s\sqrt{n\mu''} - A)\sqrt{2\pi}} \\ &= \frac{1}{(s\sqrt{n\mu''} - A)\sqrt{2\pi\mu''}} \exp\left(-A s\sqrt{n\mu''} + \frac{A^2}{2}\right) \end{aligned}$$

Collecting these results we have the following:

Theorem: If $A = \frac{8c\rho_3}{\sqrt{n}}$ is less than 1.86

$$\left. \begin{aligned} F_n(n\mu'(s)) \\ 1 - F_n(n\mu'(s)) \end{aligned} \right\} \geq \frac{1}{s\sqrt{2\pi\mu''}} e^{n(\mu(s) - s\mu'(s))} \frac{e^{-A s\sqrt{n\mu''} + \frac{A^2}{2}}}{(\sqrt{\mu''} - A/s\sqrt{n})}$$

Asymptotic Behavior of the Distribution Function

Theorem: Let n random variables have the same distribution function $F(x)$, the logarithm of the moment generating function $\mu(s)$ existing for $A < s < B$ where $A < 0 < B$. Let $F_n(x)$ be the distribution function for the sum of these random variables.

(1) If $F(x)$ is not a lattice distribution, we have asymptotically

as $n \rightarrow \infty$

$$F_n(n\mu'(s)) \sim \frac{1}{\sqrt{s|2\pi\mu''(s)|}} e^{n(\mu(s) - s\mu'(s))} \quad A < s < 0$$

$$1 - F_n(n\mu'(s)) \sim \frac{1}{\sqrt{s|2\pi\mu''(s)|}} e^{n(\mu(s) - s\mu'(s))} \quad 0 < s < B$$

(2) If $F(x)$ is a lattice distribution with maximum span h and Δ is the distance from $n\mu'(s)$ to the next lattice point in the direction away from the mean, then asymptotically as $n \rightarrow \infty$

$$F_n(n\mu'(s)) \sim \frac{e^{-|s|\Delta} h}{1 - e^{-|s|h}} \frac{1}{\sqrt{2\pi n\mu''(s)}} e^{n(\mu(s) - s\mu'(s))} \quad A < s < 0$$

$$1 - F_n(n\mu'(s)) \sim \frac{h e^{-|s|\Delta}}{1 - e^{-|s|h}} \frac{1}{\sqrt{2\pi n\mu''(s)}} e^{n(\mu(s) - s\mu'(s))} \quad 0 < s < B$$

Proof: Consider first the non-lattice case. The two results $s > 0$ and $s < 0$ are substantially the same. We prove the $s > 0$ case. As in the theorems giving upper and lower bounds, a change of variable, $y = \frac{x - n\mu'(s)}{\sqrt{n\mu''(s)}}$, reduces the problem to that of estimating the following integral

$$\int_0^{\infty} e^{-s\sqrt{n\mu''(s)}} y dH_n(y).$$

We now use the Cramér-Esseen theorem (Gnedenko and Kolmogoroff p. 210)

which states, in effect, that for any $\varepsilon > 0$ there exists n_0 such that

when $n > n_0$ we have

$$H_n(y) = \Phi(y) + \frac{e^{-y^2/2}}{\sqrt{2\pi}} \frac{\alpha_3(1-y^2)}{6\sigma^3\sqrt{n}} + B(y)$$

with $B(y) < \frac{\varepsilon}{\sqrt{n}}$. Thus the integral may be written as a sum of three integrals:

$$\int_0^\infty e^{-s\sqrt{n\mu^{11}}} y \left[d\Phi(y) + dU(y) + dB(y) \right]$$

where $U(y) = \frac{e^{-y^2/2}}{\sqrt{2\pi}} \frac{\alpha_3(1-y^2)}{6\sigma^3\sqrt{n}}$. The first integral may be evaluated exactly, on completing the square in the exponent. Its value is

$$\frac{s^2 n \mu^{11}}{e^{\frac{s^2 n \mu^{11}}{2}}} \left(1 - \Phi(s\sqrt{n\mu^{11}}) \right).$$

Using the well-known asymptotic formula for $1 - \Phi(x)$, this expression is

asymptotic to

$$\frac{s^2 n \mu^{11}}{e^{\frac{s^2 n \mu^{11}}{2}}} \frac{1}{s\sqrt{2\pi n \mu^{11}}} e^{-\frac{s^2 n \mu^{11}}{2}}$$

$$= \frac{1}{s\sqrt{2\pi n \mu^{11}}}.$$

The second integral is $o\left(\frac{1}{\sqrt{n}}\right)$. In fact, let the integral be divided into two ranges

$$\int_0^\infty = \int_0^{n^{-1/6}} + \int_{n^{-1/6}}^\infty$$

The first integral is $o\left(\frac{1}{\sqrt{n}}\right)$ because the total change in $U(y)$ in the interval is $o\left(\frac{1}{\sqrt{n}}\right)$ while the integrand is bounded. Note that $U(y)$, in addition to \sqrt{n} in the denominator, is flat at $y = 0$. Hence, as the interval of integration approaches zero, $\Delta U(y)$ is $o\left(\frac{1}{\sqrt{n}}\right)$.

The integral $\int_{n^{-1/6}}^{\infty}$ is clearly bounded by $e^{-s\sqrt{n\mu^{(1)}}} n^{-1/6} K$ where K is the total variation of $U(y)$. Since this latter is finite, and in fact even approaches zero as n increases, the term in question is certainly $o(\frac{1}{\sqrt{n}})$.

Finally, the last of the three integrals is clearly bounded by $\frac{e}{\sqrt{n}}$ and consequently is $o(\frac{1}{\sqrt{n}})$. Thus we conclude that

$$\int_0^{\infty} e^{-s\sqrt{n\mu^{(1)}}} y dH_n(y) = \frac{1}{s\sqrt{2\pi n\mu^{(1)}}} + o(\frac{1}{\sqrt{n}})$$

and, hence, that as $n \rightarrow \infty$ the tail of the original distribution with $s > 0$ has the following asymptotic formula

$$1 - F_n(\mu^{(1)}(s)) \approx \frac{1}{s\sqrt{2\pi n\mu^{(1)}}(s)} e^{n(\mu(s) - s\mu^{(1)}(s))}$$

The analysis for the case of a lattice distribution is quite similar but involves another term. We use the theorem of Esseen (Gnedenko and Kolmogoroff, p. 213) which may be phrased for our purposes as follows. For any $\varepsilon > 0$ there exists n_0 such that when $n > n_0$ we have

$$H_n(y) = \Phi(y) + \frac{e^{-y^2/2}}{\sqrt{2\pi}} \frac{\alpha_3(1-y^2)}{6\sigma^2\sqrt{n}} + \frac{e^{-y^2/2}}{\sqrt{2\pi n}} \frac{h}{\sigma} S\left(\frac{(y-\Delta)\sigma\sqrt{n}}{h}\right) + B(y)$$

with $B(y) \leq \frac{\varepsilon}{\sqrt{n}}$. In this formula σ^2 is the second moment and α_3 the third moment of the H distribution. Also, h is the maximum span, that is, the largest distance such that all jumps of the H distribution occur at multiples of this from each other. Δ is the position of the first jump in the H distribution in the positive direction. Finally, $S(Z) = [Z] - Z + \frac{1}{2}$, that is a saw-tooth function which jumps vertically from $-\frac{1}{2}$ to $+\frac{1}{2}$ at the integer values of Z and decreases linearly with slope -1 between the integers.

To estimate $\int_0^{\infty} e^{-s\sqrt{n\mu''}} y dH_n(y)$, we observe first that three of the terms are identical with those involved in the non-lattice case and consequently the integral with respect to these functions is asymptotic to $\frac{1}{s\sqrt{2\pi n\mu''}(s)}$. The only term to be evaluated is that involving the S function. This can be written as the sum of two integrals on taking the differential of the product

$$\int e^{-s\sqrt{n\mu''}} y \frac{h}{\sqrt{2\pi n\mu''}} S de^{-y^2/2} + \int e^{-s\sqrt{n\mu''}} y \frac{h}{\sqrt{2\pi n\mu''}} e^{-y^2/2} dS.$$

The first integral is $O(\frac{1}{\sqrt{n}})$. This can be seen by dividing the range of integration, 0 to $n^{-1/6}$ and $n^{-1/6}$ to ∞ , as before. The argument is essentially the same. In the first interval, the integral is small because of the flatness of $e^{-y^2/2}$ and because of the \sqrt{n} in the denominator. In the second interval, the term $e^{-s\sqrt{n\mu''}} y$ forces the integral to be small. The second integral above, integrating on dS , can be divided into an infinite sum for the jump points of S and an integral $d((y-\Delta) \circ \sqrt{n})/h$ for the sloping parts of S . The infinite sum is

$$\frac{h}{\sqrt{2\pi n\mu''}} \sum_i e^{-s\sqrt{n\mu''}} y_i e^{-y_i^2/2}$$

where the summation is over the y_i which make the argument of the S function an integer:

$$\frac{(y_i - \Delta)\sqrt{n\mu''}}{h} = K$$

where K is an integer, or

$$y_i = \frac{hK + \Delta}{\sqrt{n\mu''}}$$

Thus the sum becomes

$$\frac{h}{\sqrt{2\pi n\mu''}} \sum_K e^{-s(hK + \Delta)} e^{-(hK + \Delta)^2/2n\mu''}.$$

To estimate this sum, we use again the device of dividing the range of summation into a part from 0 to $n^{-1/6}$ and a part from $n^{-1/6}$ to ∞ . In the first of these sums, the exponential with the squared exponent approaches the constant 1 for all K in the range, and the sum reduces essentially to a geometric series. Asymptotically, then, the sum becomes

$$\frac{h}{\sqrt{2\pi n \mu^{11}}} e^{-s \Delta} \frac{1}{1 - e^{-hs}}.$$

We have still one further term to estimate, namely

$$- \int_0^{\infty} e^{-s \sqrt{n \mu^{11}}} y e^{-y^2/2} \frac{h \sqrt{\mu^{11}} \sqrt{n}}{\sqrt{2\pi n \mu^{11}}} \frac{dy}{h}$$

$$= - \int_0^{\infty} e^{-s \sqrt{n \mu^{11}}} y d\bar{\Phi}(y).$$

This term is exactly equal to the original $d\bar{\Phi}(y)$ integral and opposite in sign (since the saw-tooth slopes are in the negative direction). These two terms therefore cancel each other, and we are left with only one term of order $\frac{1}{\sqrt{n}}$. The final answer, then, is that asymptotically, with large n ,

$$\int_0^{\infty} e^{-s \sqrt{n \mu^{11}}} y dH_n(y) \approx \frac{h e^{-\Delta s}}{1 - e^{-hs}} \frac{1}{\sqrt{2\pi n \mu^{11}}}.$$

This completes the proof of the theorem.

It may be noted that if the coefficient $\frac{h e^{-\Delta s}}{1 - e^{-hs}}$ be expanded in a power series, the first terms are $\frac{1}{s} \left(1 + s \left(\frac{h}{2} - \Delta \right) + \dots \right)$. Hence, as $h \rightarrow 0$ (and also, therefore, $\Delta \rightarrow 0$) the lattice result approaches the non-lattice result, as is to be expected. It may also be seen that with $\Delta = \frac{h}{2}$ the lattice coefficient is a particularly close approximation to the non-lattice coefficient since the quantity $\frac{h}{2} - \Delta$ then vanishes. Indeed, in this case, the coefficient becomes $\frac{1}{s} \left(1 - \frac{1}{24} (h s)^2 + \dots \right)$.

Generalized Chebycheff and Chernoff Inequalities

Suppose we have a random vector (x_1, x_2, \dots, x_k) . Let $\phi(u_1, u_2, \dots, u_k)$ be everywhere non-negative and monotone increasing in all the u_i , and assume $E[\phi(x_1, x_2, \dots, x_k)]$ exists.

$$\text{Prob} [x_i \geq t_i \ (i=1, 2, \dots, k)] \leq \frac{E[\phi(x_1, x_2, \dots, x_k)]}{\phi(t_1, t_2, \dots, t_k)} \quad (1)$$

If we choose for ϕ the function

$$\phi(u_1, u_2, \dots, u_k) = e^{s_1 u_1 + s_2 u_2 + \dots + s_k u_k} \quad (\text{all } s_i \geq 0)$$

and let $\mu(s_1, s_2, \dots, s_k) = \log E(\phi) = \log$ (moment generating function of the distribution), then (1) becomes

$$\text{Prob} [x_i \geq t_i \ (i=1, 2, \dots, k)] \leq e^{\mu(s_1, s_2, \dots, s_k) - \sum s_i t_i} \quad (2)$$

This bound is minimized by choosing the s_i to satisfy

$$\frac{\partial \mu}{\partial s_i} = t_i \quad (i=1, 2, \dots, k) \quad (3)$$

If the random vectors are the sum of n independent random vectors, each with the same distribution, then the $\mu^*(s_i)$, say for the sum vector, is $n\mu(s_i)$ where $\mu(s_i)$ is the log (moment generating function) for the individual random vectors. The above result may then be translated

$$\text{Prob} [x_i^n \geq nt_i \ (i=1, 2, \dots, k)] \leq e^{n[\mu(s_1, s_2, \dots, s_k) - \sum s_i t_i]}$$

with the best choice of s_i , those which satisfy $\frac{\partial \mu}{\partial s_i} = t_i$.

Channels with Side Information at the Transmitter

Claude E. Shannon

(1)

Channels with feedback from the receiving to the transmitting point are a special case of a situation in which there is additional information available at the transmitter which may be used as an aid in the forward transmission system. In Fig. 1 the channel has an input x and an output y .

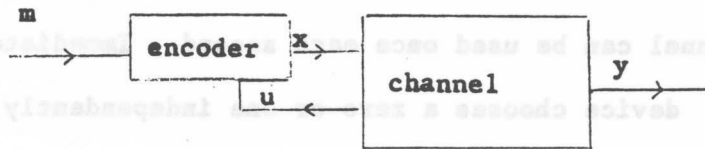


Fig. 1

There is a second output from the channel, u , available at the transmitting point, which may be used in the coding process. Thus the encoder has as inputs the message to be transmitted, m , and the side information u . The sequence of input letters x to the channel will be a function of the available part (that is, the past up to the current time) of these signals.

The signal u might be the received signal y , it might be a noisy version of this signal, or it might not relate to y but be statistically correlated with the general state of the channel. As a practical example, a transmitting station might have available a receiver for testing the current noise conditions at different frequencies. These results would be used to choose the frequency for transmission.

A simple discrete channel with side information is shown in Fig. 2

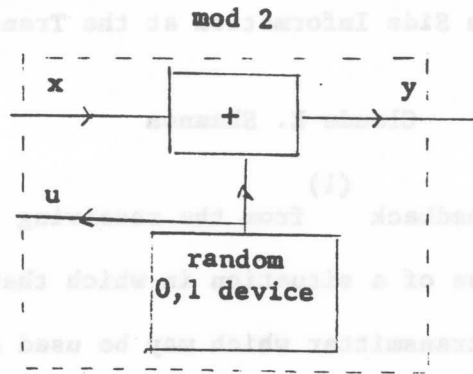


Fig. 2

In this channel, x , y and u are all binary variables; they can be either zero or one. The channel can be used once each second. Immediately after it is used the random device chooses a zero or one independently of previous choices and with probabilities $1/2$, $1/2$. This value of u then appears at the transmitting point. The next x that is sent is added in the channel modulo 2 to this value of u to give the received y . If the u side information were not available at the transmitter, the channel would be that of Fig. 3,

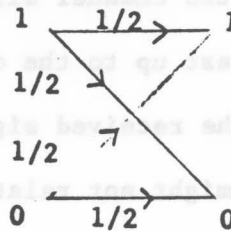


Fig. 3

a channel with capacity zero. However, with the side information available, it is possible to send one bit per second through the channel. The u information is used to compensate for the noise inside by a preliminary reversal of zero and one, as in Fig. 4.

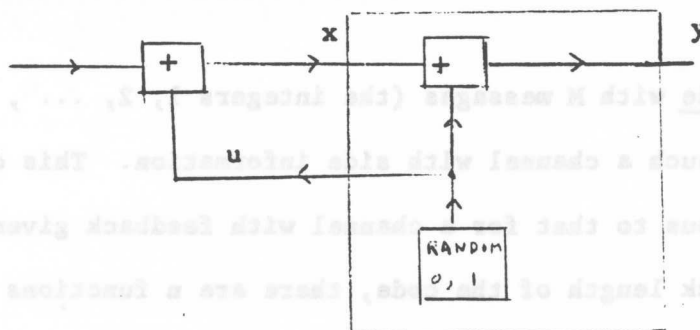


Fig. 4

Without studying the problem of side information in its fullest generality, which would involve possible historical effects in the channel, possibly infinite input and output alphabets, etc., we shall consider a moderately general case for which a simple solution has been found. See (2) also in this connection Silverman.

The memoryless discrete channel with side state information.

We consider a channel which has a finite number of possible states, s_1, s_2, \dots, s_g . At each use of the channel a new state is chosen, probability q_t for state s_t . This choice is statistically independent of previous states and previous input or output letters in the channel. The state is available as side information u at the transmitting point. When in state s_t the channel acts like a particular discrete channel K_t . Thus, its operation is defined by a set of transition probabilities $p_{ti}(j)$, $t = 1, 2, \dots, g$, $i = 1, 2, \dots, a$, $j = 1, 2, \dots, b$, where a is the number of input letters and b the number of output letters. Thus, abstractly, the channel is described by the set of state probabilities q_t and transition probabilities $p_{ti}(j)$, with q_t the probability of state t and $p_{ti}(j)$ the probability if in state t and i is transmitted, that j will be received.

A block code with M messages (the integers $1, 2, \dots, M$) may be defined as follows for such a channel with side information. This definition, incidentally, is analogous to that for a channel with feedback given previously (1).

If n is the block length of the code, there are n functions

$$f_1(m; u_1), f_2(m; u_1, u_2), f_3(m; u_1, u_2, u_3), \dots, f_n(m; u_1, u_2, \dots, u_n).$$

In these functions m ranges over the set of possible messages. Thus

$m = 1, 2, \dots, M$. The u_i all range over the possible side information

alphabet. In the particular case here each $u_i = 1, 2, \dots, g$. Each

function f_i takes values in the alphabet of input letters x of the channel.

The value $f_i(m; u_1, u_2, \dots, u_i)$ is the input x_i to be used in the code if

the message is m and the side information up to the time corresponding to i

consisted of u_1, u_2, \dots, u_i . This is the mathematical equivalent of saying

that a code consists of a way of determining, for each message m and each history of side information up to the present, the next transmitted letter.

The important feature here is that only the data available at the time i ,

namely $m; u_1, u_2, \dots, u_i$, may be used in deciding the next transmitted

letter x_i , not the side information u_{i+1}, \dots, u_n yet to appear.

A decoding system for such a code consists of a mapping or function

$h(y_1, y_2, \dots, y_n)$ of received blocks of length n into messages m ; thus h

takes values from 1 to M . It is a way of deciding on a transmitted message

given the received block y_1, y_2, \dots, y_n . For a given set of probabilities

of the messages, there will exist, for a given channel and coding and

decoding system, a calculable probability of error P_e ; the probability of

a message being encoded and received in such a way that the function h leads to deciding on a different message. We shall be concerned particularly with cases where the messages are equiprobable, each having probability $\frac{1}{M}$. The rate for such a code is $\frac{1}{n} \log M$. We are interested in the channel capacity C , that is the largest rate R such that it is possible to construct codes arbitrarily close to rate R and with probability of error P_e arbitrarily small.

It may be noted that if the state information were not available at the transmitting point, the channel would act like a memoryless channel with transition probabilities given by

$$p_i'(j) = \sum_t q_t p_{ti}(j) \quad (1)$$

Thus, the capacity C_1 under this condition could be calculated by the ordinary means for memoryless channels. On the other hand, if the state information were available both at transmitting and receiving points, it is easily shown that the capacity is then given by $C_2 = \sum_t q_t C_t$ where C_t is the capacity of the memoryless channel with transmission probabilities $p_{ti}(j)$. The situation we are interested in here is intermediate -- the state information is available at the transmitting point but not at the receiving point.

Theorem. The capacity of a memoryless discrete channel K with side state information, defined by q_t and $p_{ti}(j)$, is equal to the capacity of

the memoryless channel K' (without side information) with the same output alphabet and an input alphabet with a^g input letters $X = (x_1, x_2, \dots, x_g)$ where each $x_i = 1, 2, \dots, a$. The transition probabilities $r_X(y)$ for the channel K' are given by

$$r_X(y) = r_{x_1, x_2, \dots, x_g}(y) = \sum_t q_t p_{tx_t}(y).$$

Any code and decoding system for K' can be translated into an equivalent code and decoding system for K with the same probability of error. Any code for K has an equivocation of message (conditional entropy per letter of the message given the received sequence) at least $R - C$, where C is the capacity of K' . Thus any code with rate $R > C$ has a probability of error bounded away from zero (independent of the block length n)

$$P_e \geq \frac{R - C}{6(R + \frac{1}{n} \ln \frac{R}{C})}.$$

It may be noted that this theorem reduces the analysis of the given channel K with side information to a memoryless channel K' with more input letters but without side information. One uses known methods to determine the capacity of this derived channel and this gives the capacity of the original channel. Furthermore, codes for the derived channel may be translated into codes for the original channel with identical probability of error. (Indeed, all statistical properties of the codes are identical.)

We first show how codes for K' may be translated into codes for K . A code word for the derived channel K' consists of a sequence of n letters X from the X input alphabet of K' . A particular input letter X of this channel

may be recognized as a particular function from the state alphabet to the input alphabet x of channel K . The full possible alphabet of X consists of the full set of a^g different possible functions from the state alphabet with g values to the input value with a values. Thus, each letter $X = (x_1, x_2, \dots, x_g)$ of a code word for K' may be interpreted as a function from state u to input alphabet x . The translation of codes consists merely of using the input x given by this function of the state variable. Thus if the state variable u has the value 1, then x_1 is used in channel K ; if it were state k , then x_k . In other words, the translation is a simple letter by letter translation without memory effects depending on previous states.

The codes for K' are really just another way of describing certain of the codes for K -- namely those where the next input letter x is a function only of the message m and the current state u , and does not depend on the previous states.

It might be pointed out also that a symple physical device could be constructed which, placed ahead of the channel K , makes it look like K' . This device would have the X alphabet for one input and the state alphabet for another (this input connected to the u line of Fig. 1). Its output would range over the x alphabet and be connected to the x line of Fig. 1.

Its operation would be to give an x output corresponding to the X function of the state u . It is clear that the statistical situations for K and K' with the translated code are identical. The probability of an input word for K' being received as a particular output word is the same as that for the corresponding operation with K . This gives the first part of the theorem.

To prove the second part of the theorem, we will show that in the channel K , the change in conditional entropy (equivocation) of the message m at the receiving point when a letter is received cannot exceed C (the capacity of the channel K'). In Fig. 1, we let m be the message; x , y , u be the next input letter, output letter and state letter. Let U be the past sequence of u states from the beginning of the block code to the present (just before u), and Y the past sequence of output letters up to the current y . We are assuming here a given block code for encoding messages. The messages are chosen from a set with certain probabilities (not necessarily equal). Given the statistics of the message source, the coding system, and the statistics of the channel, these various entities m , x , y , U , Y all belong to a probability space and the various probabilities involved in the following calculation are meaningful. Thus the equivocation of message when Y has been received, $H(m|Y)$, is given by

$$\begin{aligned} H(m|Y) &= - \sum_{m,Y} P(m,Y) \log P(m|Y) \\ &= - \langle \log P(m|Y) \rangle \end{aligned}$$

(The symbol $\langle G \rangle$ here and later means the average of G over the probability space.) The change in equivocation when the next letter y is received by

$$\begin{aligned} H(m|Y) - H(m|Y,y) &= - \langle \log P(m|Y) \rangle + \langle \log P(m|Y,y) \rangle \\ &= \left\langle \log \frac{P(m|Y,y)}{P(m|Y)} \right\rangle \end{aligned}$$

$$\begin{aligned}
&= \left\langle \log \frac{P(m, Y, y) P(Y)}{P(Y, y) P(m, Y)} \right\rangle \\
&= \left\langle \log \frac{P(y|mY) P(Y)}{P(Y, y)} \right\rangle \\
&= \left\langle \log \frac{P(y|mY)}{P(y)} \right\rangle - \left\langle \log \frac{P(Y, y)}{P(Y) P(y)} \right\rangle
\end{aligned}$$

$$H(m|Y) - H(m|Y, y) \leq \left\langle \log \frac{P(y|mY)}{P(y)} \right\rangle \quad (1)$$

The last reduction is true since the term $\left\langle \log \frac{P(Y, y)}{P(Y) P(y)} \right\rangle$ is an average mutual information and therefore non-negative. Now note that by the independency requirements of our original system

$$P(y|x) = P(y|x, m, u, U) = P(y|x, m, u, U, Y)$$

Now since x is a strict function of m, u , and U (by the coding system function) we may omit this in the conditioning variables

$$P(y|m, u, U) = P(y|m, u, U, Y)$$

$$\frac{P(y, m, u, U)}{P(m, u, U)} = \frac{P(y, m, u, U, Y)}{P(m, u, U, Y)}$$

Since the new state u is independent of the past $P(m, u, U) = P(u)P(m, U)$ and $P(m, u, U, Y) = P(u) P(m, U, Y)$. Substituting and simplifying

$$P(y, u|m, U) = P(y, u|m, U, Y)$$

Summing on u gives

$$P(y|m, U) = P(y|m, U, Y)$$

Hence:

$$H(y|m,U) = H(y|m,U,Y) \leq H(y|m,Y) \\ - \langle \log P(y|m,U) \rangle \leq - \langle \log P(y|m,Y) \rangle$$

Using this in (1)

$$H(m|Y) - H(m|Y,y) \leq \left\langle \log \frac{P(y|m,U)}{P(y)} \right\rangle \quad (2)$$

We now wish to show that $P(y|m,U) = P(y|X)$. Here X is a random variable specifying the function from u to x imposed by the encoding operation for the next input x to the channel. Equivalently, X corresponds to an input letter in the derived channel K' . We have

$P(y|x,u) = P(y|x,u,m,U)$. Furthermore, the coding system used implies

a functional relation for determining the next input letter x , given

m, U and u . Thus $x = f(m,U,u)$. If $f(m,U,u) = f(m',U',u)$ for two

particular pairs (m,U) and (m',U') but for all u , then it follows

that $P(y|m,U,u) = P(y|m',U',u)$ for all u and y ; since m, U and u

lead to the same x as m', U' , and u . From this we obtain

$$P(y|m,U) = \sum_u P(u)P(y|m,U,u) = \sum_u P(u)P(y|m',U',u) = P(y|m',U').$$

In other words, (m,U) pairs which give the same function $f(m,U,u)$

give the same value of $P(y|m,U)$ or, said another way, $P(y|m,U) = P(y|X)$.

Returning now to our inequality (2), we have

$$H(m|Y) - H(m|Y,y) \leq \left\langle \log \frac{P(y|X)}{P(y)} \right\rangle$$

$$\leq \max_{P(X)} \left\langle \log \frac{P(y|X)}{P(y)} \right\rangle$$

$$H(m|Y) - H(m|Y,y) \leq C.$$

equivocation

This is the desired inequality on the equivocation. The/ cannot be reduced by more than C , the capacity of the derived channel K' for each received letter. In particular in a block code with M equiprobable messages, $R = \frac{1}{n} \log M$, If $R > C$, then at the end of the block the equivocation must still be at least $nR - nC$, since it starts at nR and can only reduce at most C for each of the n letters.

It is known that if the equivocation per letter is at least $R - C$ then the probability of error in decoding is at least

$$P_e \geq \frac{R - C}{6(R + \frac{1}{n} \log \frac{R}{C})}$$

Thus the probability of error is founded away from zero regardless of the block length n , if the code attempts to send at a rate $R > C$.

This concludes the proof of the theorem.

As an example of this theorem, consider a channel with two output letters, any number a of input letters and any number g of states. Then the derived channel K' has two output letters and a^g input letters. However, in a channel with just two output letters, only two of the input letters need be used to achieve channel capacity, as shown in (3). Namely, we should use in K' only the two letters with maximum and minimum transition probabilities to one of the output letters. These two may be found as follows. The transition probabilities for a particular letter of K' are averages of the corresponding transitions for a set of letters for K , one for each state. To maximize the transition probability to one of the output letters, it is clear that we should choose in each state the letter with the maximum transition to that output letter. Similarly, to

minimize, one chooses in each state the letter with the minimum transition probability to that letter. These two resulting letters in K' are the only ones used, and the corresponding channel gives the desired channel capacity. Formally, then, if the given channel has probabilities $p_{ti}(1)$ in state t for input letter i to output letter 1, and $p_{ti}(2) = 1 - p_{ti}(1)$ to the other output letter 2, we calculate;

$$p_1 = \sum_t q_t \max_i p_{ti}(1)$$

$$p_2 = \sum_t q_t \min_i p_{ti}(1)$$

The channel K' with two input letters having transition probabilities p_1 and $1 - p_1$ and p_2 , $1 - p_2$ to the two output letters respectively, has the channel capacity of the original channel K .

Another example, with three output letters, two input letters and three states, is the following. The probability matrices for the three states are: (the states assumed to each have probability $1/3$)

	State 1			State 2			State 3		
	1	0	0	0	1	0	0	0	1
	0	1/2	1/2	1/2	0	1/2	1/2	1/2	0

In this case there are $2^3 = 8$ input letters in the derived channel K' .

The matrix of these is as follows:

$1/2$	$1/2$	0
0	$1/2$	$1/2$
$1/2$	0	$1/2$
$2/3$	$1/6$	$1/6$
$1/6$	$2/3$	$1/6$
$1/6$	$1/6$	$2/3$
$1/3$	$1/3$	$1/3$
$1/3$	$1/3$	$1/3$

If there are only three output letters one need use only three input letters to achieve channel capacity, and in this case it is readily shown that the first three can (and in fact must) be used. Due to the symmetry, these three letters must be used with equal probability and the resulting channel capacity is $\log 3/2$.

In the original channel, it is easily seen that, if the state information were not available, the channel would act like one with the transition matrix

$1/3$	$1/3$	$1/3$
$1/3$	$1/3$	$1/3$

This channel clearly has zero capacity. On the other hand, if the state information were available at the receiving point or at both the receiving point and the transmitting point, the two input letters can be perfectly distinguished and the channel capacity is $\log 2$.

Some Miscellaneous Results in Coding Theory

Claude E. Shannon

This paper contains a number of somewhat miscellaneous results centered chiefly on the problem of coding sources into noiseless channels, including cases where the channel symbols have different durations or costs.

The number of sequences of a given length

Suppose a number of letters are available whose lengths (or durations) are a_1, a_2, \dots, a_g and we wish a bound on the number of sequences of total length l . Here it is assumed that any sequence of letters is allowed. We define $N(l)$ to be the number of different sequences whose total length is greater than $l - a_{\min}$ but not greater than l . Here a_{\min} is the smallest a_i . Thus $N(l)$ might be thought of as the number of sequences of length l where we allow filling out with a blank to an extent up to the shortest letter. This definition makes $N(l)$ better behaved (e.g., it is now monotone increasing) than if we count only sequences of exactly length l .

$N(l)$ satisfies the difference equation

$$N(l) = N(l - a_1) + N(l - a_2) + \dots + N(l - a_g) \quad l > 0$$

as we see by noting that each sequence of length l must end in one or another of the available letters. Furthermore, the boundary conditions may be taken

to be $N(l) = 0$ for $l < 0$ and $N(l) = 1$ for $0 \leq l < a_{\min}$.

Associated with the difference equation is the following characteristic equation:

$$1 = X^{-a_1} + X^{-a_2} + \dots + X^{-a_g}$$

Since all the a_i are positive and real, the right-hand member is a strictly monotone decreasing function of X and varies from ∞ to 0 when X goes from 0 to ∞ . Consequently, the characteristic equation has a unique positive real root W .

Theorem: For all l , $N(l) \leq W^l$. For all $l \geq 0$, $N(l) \geq W^{l-a_{\max}}$.

This will be proved by a kind of induction on increasing intervals of l , each interval of length a_{\min} . Consider first the upper bound W^l . This is certainly true for $0 \leq l \leq a_{\min}$, since in this range $N(l) = 1$ and $W > 1$. Now assume the upper bound true out to some l_1 . Then for l in the range $l_1 \leq l \leq l_1 + a_{\min}$ we have

$$\begin{aligned} N(l) &= N(l - a_1) + N(l - a_2) + \dots + N(l - a_g) \\ &\leq W^{l-a_1} + W^{l-a_2} + \dots + W^{l-a_g} \\ &= W^l \end{aligned}$$

Thus the theorem is then true for the increased interval up to $l_1 + a_{\min}$. It follows that the bound is true for all l .

The lower bound is very similar. It is certainly true for $0 \leq l \leq a_{\max}$ since $N(l) \geq 1$ in this range and $W^{l-a_{\max}} < 1$. The inductive step goes through as before. Assuming that for $0 \leq l \leq l_1$ (with $l_1 \geq a_{\max}$) we have $N(l) \geq W^{l-a_{\max}}$, then in the extended range from l_1 to $l_1 + a_{\min}$ we have

$$N(l) = \sum N(l - a_i)$$

$$\geq \sum W^{l - a_i - a_{\max}} \\ = W^{l - a_{\max}}$$

Thus by extending the range with steps of a_{\min} we obtain the result for all positive l .

This result, of course, relates to how rapidly it is possible to approach the capacity of a noiseless channel with unequal symbol lengths. Thus for $l \geq 0$, from this theorem

$$(\log W) - \frac{a_{\max}}{l} \leq \frac{1}{l} \log N(l) \leq \log W$$

The approach of possible signalling rate to the capacity $\log W$ is rapid, the discrepancy at most $\frac{a_{\max}}{l}$.

An interesting alternative proof that $N(l) \leq W^l$ can be given as follows. Assume, in contradiction, that for some l , $N(l) > W^l$. Then, since $N(0) \leq W^0$, there is a greatest lower bound of l 's, say l^* , for which the theorem fails. In the interval $l^* \leq l \leq l^* + \frac{1}{2} a_{\min}$ there must be an l , say l_1 , for which the theorem fails. Subdivide the sequences of length l_1 into subsets according to the first letter. Let the fractional number in the subset beginning with the letter i be f_i ($i = 1, 2, \dots, g$). Choose the subset for which $a_i^{-1} \log f_i^{-1}$ is a minimum. In a sense, this means the subset which conveys the least information, $\log f_i^{-1}$, per unit time in its first letter. The minimum value of $a_i^{-1} \log f_i^{-1}$ among the different subsets is less than or equal to $\log W$. To see this, suppose, in contradiction, that for all i , $a_i^{-1} \log f_i^{-1} > \log W$. Then $f_i < W^{-a_i}$ and, summing on i , $1 = \sum f_i < \sum W^{-a_i} = 1$, a contradiction. Hence the subset

chosen will have $a_i^{-1} \log f_i^{-1} \leq \log W$, or $f_i \geq W^{-a_i}$. If we delete the first letter from all sequences in this subset, we are left with a set of more than $W^{\ell_1 - a_i}$ sequences of length $\ell_1 - a_i$. Thus $N(\ell_1 - a_i) > W^{\ell_1 - a_i}$. Since $\ell_1 - a_i < \ell^*$, this contradicts the assumption that ℓ^* was the greatest lower bound of ℓ 's for which the theorem fails. Hence the theorem is true for all ℓ .

The case with unequal letters and a finite set of constraints

A more general problem of the same sort relates to sequences which are subject to a finite state set of constraints. Thus, suppose there are d states and that in state i , letters of lengths $\ell_{\alpha ij}$ are permitted leading to state j . The index α ranges over the different letters going from state i to state j and j ranges over the different states which can follow state i . Now let $N_{ij}(\ell)$ be the number of sequences which are possible and which start in state i , end in state j and are of length ℓ . These quantities are readily seen to satisfy the difference equations

$$N_{ij}(\ell) = \sum_{\alpha, k} N_{ik}(\ell - \ell_{\alpha kj}) \quad \ell > 0 \quad (1)$$

$$N_{ij}(\ell) = 0 \quad \ell < 0$$

The corresponding characteristic equations are

$$A_j = \sum_{\alpha, i} A_i W^{-\ell_{\alpha ij}} \quad (2)$$

Let W be the largest real root (there is a positive real root as shown in the appendix) of the determinant equation:

$$\left| \sum_{\alpha} W^{-\ell_{\alpha ij}} - \delta_{ij} \right| = 0$$

and let A_i be a corresponding (positive) solution of (2). We will assume the graph of the constraints has complete accessibility so it is possible to go from any state to any other. Then all the A_i are positive (none vanish).

We will show that the number of sequences of length ℓ , starting in state i and ending in j , $N_{ij}(\ell)$, is bounded by

$$N_{ij}(\ell) \leq \frac{A_j}{A_i} W^\ell$$

This is certainly true for $\ell < 0$ and also for $\ell = 0$ since then both sides are one if $i = j$, and otherwise the left side is zero with the right positive. We now proceed by the inductive type process as before, assuming the inequality out to some ℓ_1 and then show it follows for ℓ out to ℓ_1 plus the minimum ℓ_{cij} .

$$\begin{aligned} N_{ij}(\ell) &= \sum_{cs} N_{is}(\ell - \ell_{csj}) \\ &\leq \sum_{cs} \frac{A_s}{A_i} W^{\ell - \ell_{csj}} \quad \ell \leq \ell_1 + \min \ell_{cij} \\ &= \frac{W^\ell}{A_i} \sum_{cs} A_s W^{-\ell_{csj}} \\ &= W^\ell \frac{A_j}{A_i} \end{aligned}$$

Thus the inductive step carries the inequality up to $\ell = \ell_1 + \min \ell_{cij}$ and hence it is true for all ℓ .

An explicit code for a variable length alphabet

It is possible to generalize a coding process we have described elsewhere for a binary alphabet to the case where there are a number of symbols of different "durations" or, more generally, with certain associated "costs." It is desired to encode a finite set of possible messages with associated probabilities p_1, p_2, \dots, p_n into sequences of letters chosen from an alphabet where the letter i has cost or duration ℓ_i and it is desired in the code to minimize the expected cost. This problem has been studied ^{by Allen Newman} in a thesis by Richard Harmons.

We shall use in our analysis a curious notation for real numbers based on unequal values for various digits. In the ordinary decimal notation, the range from 0 to 1 is divided into ten equal intervals. These are labeled with the digits from 0 to 9. Each of these intervals is again subdivided equally and again given labels. In the notation system we are now describing, the interval is subdivided into arbitrary sub-intervals of length $\lambda_0, \lambda_1, \dots, \lambda_{n-1}$, not necessarily equal but with $\sum \lambda_i = 1$. If a real number between 0 and 1 falls in the interval λ_k (closed on the left, open on the right) its first digit is k . All of the intervals are subdivided in the same proportions and this determines the second digit, etc.

This notation system has many of the properties of ordinary binary, ternary, etc., systems such as unicity of representation, apart from numbers terminating in an infinite sequence of 0's or $(n-1)$'s. However, it does differ in certain important respects. For example, if a real number is chosen at random, then in an ordinary decimal notation we

expect one-tenth of each value of digit. In this notation we expect λ_i of digit i .

Returning now to the coding problem, we recall that if a set of channel letters have durations $\ell_1, \ell_2, \dots, \ell_n$ the corresponding channel capacity is $C = \log W_0$ where W_0 is the unique positive real root of

$$\sum_i W^{-\ell_i} = 1$$

Given a set of ℓ_i and the corresponding W_0 we define a subdivision of the unit interval and a corresponding notation for real numbers by the quantities

$$\lambda_0 = W_0^{-\ell_1}, \lambda_1 = W_0^{-\ell_2}, \dots, \lambda_{n-1} = W_0^{-\ell_n}$$

Since these are all positive and their sum is unity they form a satisfactory subdivision.

Now let a set of messages have probabilities $p_1 \geq p_2 \geq \dots \geq p_m$ and let $\sum_{i=1}^k p_i = P_k$ so P_k is the cumulative probability for the first k when the messages are arranged in order of decreasing probability.

The code to be used is defined as follows. Let P_k be expanded in the notation defined by the subdivision $W_0^{-\ell_i}$ out to just enough places to make the uncertainty due to "digits" beyond this point less than p_k .

In other words, if P_k is represented in this notation system by the sequence $a_{k1}, a_{k2}, a_{k3}, \dots$ then we carry out the expansion for P_k to t places where t is chosen to make

$$W^{-\sum_{i=1}^t \ell_{a_{ki}}} < p_k \leq W^{-\sum_{i=1}^{t-1} \ell_{a_{ki}}} \quad (1)$$

The code we are defining represents message k by the sequence of channel symbols corresponding to the t digits of this expansion of P_k . It should first be noted that this does in fact form a reversible code. It satisfies the so-called prefix condition — no code word is the beginning of any other code word. Indeed the code word corresponding to P_k defines an interval including P_k and of width less than p_k . This interval consequently does not include P_{k-1} or any earlier P_i and the code word must differ in some "digit" from all preceding code words. Consequently all code words differ and any sequence of code words is uniquely decipherable.

We now wish to estimate the expected length of code words, that is,

$$\sum_{k,i} p_k \ell_{a_{ki}}. \text{ From (1) we have}$$

$$\log W \cdot \sum_{i=1}^t \ell_{a_{ki}} \geq \log p_k^{-1} \geq \log W \cdot \sum_{i=1}^{t-1} \ell_{a_{ki}}$$

Multiplying by p_k and summing over all k gives

$$\log W \cdot \sum_k p_k L_k \geq \sum_k p_k \log p_k^{-1} \geq \log W \cdot \sum_k p_k (L_k - \ell_{\max})$$

where $L_k = \sum_{i=1}^t \ell_{a_{ki}}$ is the length of the code word for message k and, on the right, we have underestimated by replacing the last term in the sum on i by its maximum possible value, ℓ_{\max} , the largest duration of any letter.

Now recognizing that $\sum_k p_k \log p_k^{-1} = H$, the entropy or average information

of the message source, and using $L = \sum p_k L_k$ to represent the expected length of a code word this may be written

$$\bar{L} \log W \geq H \geq (\bar{L} - l_{\max}) \log W$$

or

$$\frac{H}{\log W} \leq L \leq \frac{H}{\log W} + l_{\max}$$

This is our desired result. Of course the lower bound holds for any reversible code. The upper bound shows that one can approximate the ideal lower bound to within l_{\max} . In particular, if one is working with messages which consist of blocks or n -grams of text, then H becomes nH_n where H_n is the entropy per letter for blocks of length n . As n increases, H_n approaches H , the entropy per letter of the message source.

Dividing the inequalities by n we have, in this case,

$$\frac{H_n}{\log W} \leq \frac{L}{n} \leq \frac{H_n}{\log W} + \frac{l_{\max}}{n}$$

In other words, the average code length per letter of message has a discrepancy $\frac{l_{\max}}{n}$ at most from its ideal value on the basis of n -gram entropy. This is closely analogous to our previous result with channel letters of equal duration.

An inequality for a Huffman-type code

A Huffman code (2) for cases of equal cost binary symbols is optimal in giving the minimum expected length and must therefore have an expected length less than or equal to $H + 1$ since, as shown above,

or in (1), these digit expansion codes which are not necessarily optimal, satisfy this inequality. Peter Elias suggested the desirability of a direct proof from the Huffman procedure of this upper bound. In solving this problem a slightly stronger result was obtained as follows.

Theorem: In a Huffman binary code, the expected word length \bar{L} satisfies $H \leq \bar{L} \leq H + 1 - 2p_{\min}$, where H is the entropy (in bits) of the set of probabilities and p_{\min} is the smallest probability in the set.

Proof: The lower bound is of course well known. The upper bound will be proved by induction. We will assume it true for all codes corresponding to trees with $n - 1$ branch points and show that it follows (Fig. 1) for those with n branch points. Consider, then, a Huffman tree with n branch points and focus attention on the two smallest probabilities. These occur, by the method of construction, at ends of one fork. Let these probabilities be p and q with, say, $p \leq q$. If we delete the pq branches leaving $P = p + q$ at the junction, we have left a Huffman tree (because of the method of construction) with $n - 1$ junctions and to which our inductive assumption applies. Let unprimed letters relate to this tree and primed letters to the enlarged tree. Then we have $p = p'_{\min}$ (the minimum probability for the enlarged tree) and since both p and q are less than or equal to p_{\min} (the min probability for the smaller tree), $p \leq p_{\min}$. Also the average code lengths \bar{L} and \bar{L}' and the entropy H and H' are clearly related as follows:

$$\bar{L}' = \bar{L} + P$$

$$H' = H + P H\left(\frac{p}{P}, \frac{q}{P}\right)$$

Finally by inductive assumption

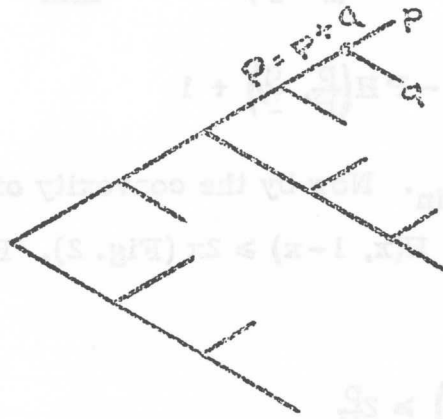


Fig. 1

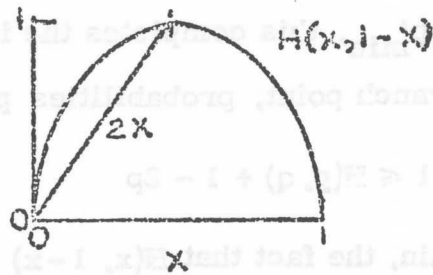


Fig. 2

$$\bar{l} \leq H + 1 - 2p_{\min}$$

$$\text{Hence } \bar{l}' \leq H + 1 - 2p_{\min} + P$$

$$= H' - PH\left(\frac{p}{P}, \frac{q}{P}\right) + 1 - 2p_{\min} + P$$

$$\leq H' - PH\left(\frac{p}{P}, \frac{q}{P}\right) + 1 \quad (2)$$

since $P \leq 2p_{\min}$. Now by the convexity of the curve $H(x, 1-x)$ we have, for $x \leq \frac{1}{2}$, that $H(x, 1-x) \geq 2x$ (Fig. 2). Hence, recalling that $p \leq q$ we have

$$H\left(\frac{p}{P}, \frac{q}{P}\right) \geq \frac{2p}{P}$$

$$PH\left(\frac{p}{P}, \frac{q}{P}\right) \geq 2p$$

Using this in the above inequality (2) we conclude

$$\bar{l}' \leq H' + 1 - 2p$$

Since $p = p'_{\min}$, this completes the induction. The theorem is true for one branch point, probabilities p and $q \geq p$, since in this case

$$\bar{l} = 1 \leq H(p, q) + 1 - 2p$$

using, again, the fact that $H(x, 1-x) \geq 2x$.

This result is easily generalized to the case where there are b available (equal length) letters in the alphabet. In this case it can be shown that $\bar{l} \leq \frac{H}{\log b} + 1 + d p_{\min}$ where d is the number of branches on the minimum probability branchpoint of the tree. Thus d is the remainder if $b - 1$ is subtracted from n , the number of messages

enough times to give a remainder less than or equal to b . The proof of this result is by the obvious generalization of the above proof and is left to the reader.

Appendix: Existence of the characteristic equation root

Lemma: Given $f_{ij}(\omega)$ ($i, j = 1, 2, \dots, d$) continuous functions of ω in the range $a \leq \omega \leq b$ and in this range $f_{ij}(\omega) \geq 0$, $\sum_j f_{ij}(\omega) > 0$, $f_{ij}(a) < \frac{1}{d}$, $f_{ij}(b) > d$, then there exists W , $a \leq W \leq b$ and a set of $X_i \geq 0$, $\sum X_i = 1$, such that

$$|f_{ij}(W) - \delta_{ij}| = 0$$

$$\sum_i X_i f_{ij}(W) = X_j$$

Proof: Consider the $(d+1)$ dimensional region R whose points are (X_1, \dots, X_d, W) , where $X_i \geq 0$, $\sum X_i = 1$, $a \leq W \leq b$. This is a topological image of a sphere and its interior. For W in the range from a to b , consider the continuous mapping

$$X_j \rightarrow Y_j = \frac{\sum_i X_i f_{ij}(W)}{\sum_{ij} X_i f_{ij}(W)} \quad W \rightarrow V = \begin{cases} V_1 = W + 1 - \sum_{ij} f_{ij}(W) X_i & \text{if } a \leq V_1 \leq b \\ a & \text{if } V_1 < a \\ b & \text{if } V_1 > b \end{cases}$$

Note that the denominator for Y_j does not vanish because of our assumption that $\sum_j f_{ij}(X) > 0$ and hence the Y_j are well defined. Also the Y_j are non-negative and $\sum Y_j = 1$. Finally $a \leq V \leq b$. Hence

this maps points (X_i, W) in R continuously into points (Y_i, V) in R .

Consequently, by the Brouwer fixed point theorem there exists a point (X_i, W) which is mapped into itself, that is, a point for which

$\sum_i X_i f_{ij}(W) = X_j \sum_{ij} X_i f_{ij}(W)$, $W = V$. The value of W for the fixpoint clearly is not a or b since these points are moved upward or downward by our assumptions. Hence for the fixpoint we have

$W = W + 1 - \sum_{ij} f_{ij}(W) X_i$ or $\sum_{ij} f_{ij}(W) X_i = 1$. It follows that for the fixpoint

$$\sum_{ij} f_{ij}(W) X_i = X_j$$

$$|f_{ij}(W) - \delta_{ij}| = 0$$

Let the elements a_{ij} of a matrix be non-negative. Suppose there is an eigen vector A_i all of whose components are positive, $A_i > 0$, and the corresponding characteristic value is λ_0 . We will show that for any other characteristic value λ_1 we have $|\lambda_1| \leq \lambda_0$. Let B_i be a characteristic vector for λ_1 where we adjust the length of this vector as follows. Choose its length in such a way that $A_i - |B_i| \geq 0$ for all i and the equality holds for at least one i , say $i = h$, so that $A_h = |B_h|$. It is clear that this can be done since with zero length all components of B are less than those of A and increasing continuously, eventually a first one of the $|B_i|$ reaches its corresponding A_i . We now have

$$\sum_i A_i a_{ij} = \lambda_0 A_j \quad (1)$$

$$\sum_i B_i a_{ij} = \lambda_1 B_j \quad (2)$$

$$\sum_i |B_i| a_{ij} \geq |\lambda_1| |B_j| \quad (3)$$

Subtracting these equations for $j = h$

$$\begin{aligned} \sum_i (A_i - |B_i|) a_{ih} &\leq \lambda_0 A_h - |\lambda_1| |B_h| \\ &= (\lambda_0 - |\lambda_1|) A_h \end{aligned} \quad (4)$$

All terms in the sum at the left are non-negative and also A_h is definitely positive. It follows that $\lambda_0 - |\lambda_1| \geq 0$.

Error Probability Bounds for Noisy Channels

C. E. Shannon

This paper gives a simplified proof akin to that published previously⁽¹⁾ but leading to tighter bounds on the error probability and to a simpler final result. We consider a discrete memoryless channel defined by a set of letter transition probabilities $p_j(i)$. Assume a given assignment of input letter probabilities P_i . These might be, but not necessarily, the set which gives channel capacity. Let $Q_j = \sum_i P_i p_j(i)$ be the output letter probabilities that would result if the P_i were used for input probabilities.

We consider, as usual, a random code ensemble based on the P_i containing $M = e^{nR}$ messages each with code word of length n . In the ensemble of codes, M messages, say the integers from 1 to M , are mapped independently into the possible input words of length n for the channel. A message is mapped into a code word with probability equal to that of the code word produced by the product probabilities generated by the P_i . Thus, the various possible codes in the ensemble have associated probabilities equal to the probability of their occurrence under this system. We wish to overbound the average error probability for this ensemble of codes with a decoding system to be described, where the error probabilities of individual codes are weighted with the probabilities associated with the particular codes.

The mutual information $I(u; v)$ between an input word u and an output word v is given by

$$I(u; v) = \log \frac{Pr_1(v|u)}{Pr_1(v)}$$

where Pr_1 means probability calculated by the given letter assignment P_1 and the given transitions $p_i(j)$, (extended independently to blocks of length n). $I(u; v)$ may be thought of here as a number associated with any input word-output word pair. $I(u; v)$ is the sum of the mutual informations between corresponding letters of u and v . Thus if u consists of the letters u_1, u_2, \dots, u_n and v of v_1, v_2, \dots, v_n then, because of the independence of channel and letter assignments, we have

$$\begin{aligned} I(u; v) &= \log \frac{\prod_i Pr_1(v_i|u_i)}{\prod_i Pr_1(v_i)} = \sum_i \log \frac{Pr_1(v_i|u_i)}{Pr_1(v_i)} \\ &= \sum_i I(u_i; v_i) \end{aligned}$$

If we now think of choosing an input word u and an output word v according to some joint probability, then $I(u; v)$ becomes a random variable. In particular, we may choose an input word u according to the product probability measure obtained from the probability assignments P_i , and then an output word v according to the transition probabilities $p_i(j)$, (independently applied to the letters of u). In the ensemble of random codes, input words u and noisy received words v will occur with this joint probability measure.

We define a decoding system for codes in the ensemble as follows. Any received word v is decoded as that message in the code in question whose code word u has the largest $I(u, v)$. (If several have equal values,

take the smallest numbered message from this set.) This might be called maximum information decoding. It must be remembered, however, that mutual information is here calculated by the original probability assignments produced by the P_1 . It is not necessarily maximum information decoding for any particular code or word in a code in the ensemble. It is actually, however, equivalent to decoding as the most probable cause of the received word, and therefore is optimal to give small error probability. This is because all messages have equal a priori probability so if $\log \frac{P(v|u_1)}{P(v)} > \log \frac{P(v|u_2)}{P(v)}$ then $\frac{P(v|u_1)}{P(v)} > \frac{P(v|u_2)}{P(v)}$. Hence if message m_1 is mapped into u_1 and m_2 into u_2 it follows from their equal prior probability that $P(m_1|v) > P(m_2|v)$.

We may also define a second joint probability measure for (u, v) pairs as follows. Consider choosing a u word according to the assigned probabilities and a v word independently according to the assigned probabilities. This joint probability measure we denote by Pr_2 . We may also consider $I(u, v)$ as a random variable with this set of probabilities $Pr_2(u, v)$ for (u, v) pairs. However, a peculiar point arises in that some of the $P(v|u)$ may be zero. For these (u, v) pairs, $I(u, v) = \log \frac{P(v|u)}{P(v)}$ is undefined. (It approaches $-\infty$ as $P(v|u)$ approaches zero.) This caused no trouble in the Pr_1 probability measure since these (u, v) pairs had zero probability in that case. Here, however, these (u, v) pairs may have positive probability. We may still, however, consider the distribution function for $I(u, v)$ in the new Pr_2 measure. Thus $Pr_2[I \geq a]$ means the sum of probabilities of all (u, v) pairs in this measure

for which $I(u, v)$ is defined and at least α . In other words, calculate the distribution function as though there were a certain probability of I being at $-\infty$. The cumulative distribution function from the left would start not at zero but at a positive value if there were some (u, v) pairs with $P(v|u) = 0$.

Lemma: For any α , the average error probability P_e for the described ensemble of codes is bounded by

$$P_e \leq \Pr_1[I \leq \alpha] + M \Pr_2[I \geq \alpha]$$

Proof: In the ensemble of codes, input words and received versions of those occur with the probability measure $\Pr_1(u, v)$. Thus, in the lemma, the term $\Pr_1[I \leq \alpha]$ can be identified with the probability of a message resulting in a received word with mutual information as low or lower than a threshold level α .

The term $\Pr_2[I \geq \alpha]$ may be interpreted as follows. In the ensemble of codes, suppose message number 1 occurs. This will give rise to various received v 's with probability (over the ensemble) given by $\Pr_1(v)$. This is because message 1 is mapped into all possible u 's with the assigned u probabilities. Now consider the probability that message number 2 has a mutual information with the received version of message 1 greater than or equal to α . This is given by $\Pr_2[I \geq \alpha]$, since in the ensemble message 2 is mapped independently into the u space. The same applies, of course, to messages 3, 4, ..., M and, in fact, for all messages other than the actual cause of the received word. The probability that any message (apart from the actual cause) has a mutual information

with the received v exceeding α is given exactly by

$$1 - (1 - \Pr_2[I \geq \alpha])^{M-1} \leq (M-1) \Pr_2[I \geq \alpha] \\ < M \Pr_2[I \geq \alpha]$$

Thus the probability that either the actual message has a mutual information less than α or that some other message has a greater than α mutual information with the received v is bounded by

$$\Pr_1[I \leq \alpha] + M \Pr_2[I \geq \alpha]$$

(The probability of either or both of two events can always be bounded by the sum of their individual probabilities, whether or not the events are independent.) If neither of these events occurs, the decoding will be correct since the mutual information with the actual cause is greater than α and that with all other messages is less than α . Thus the error probability in the ensemble is bounded by

$$P_e \leq \Pr_1[I \leq \alpha] + M \Pr_2[I \geq \alpha]$$

As an example of a random code ensemble, consider the following situation. Suppose there are two input words and two output words with the transition probabilities shown in Fig. 1.

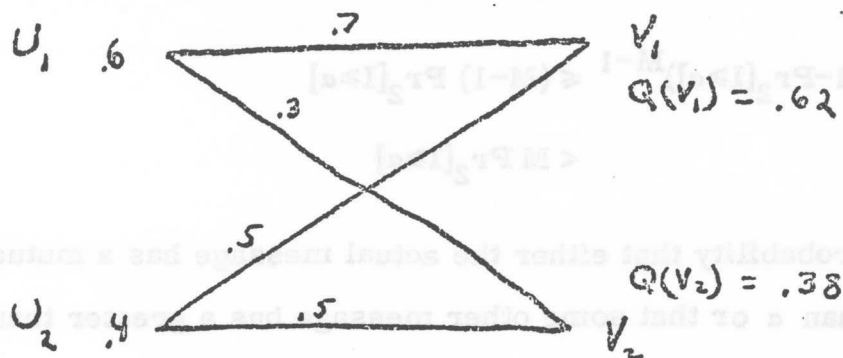


FIG. 1

The arbitrary assignments of probability .6 and .4 have been made to U_1 and U_2 , and this results in $.7 \times .6 + .5 \times .4 = .62$ for $Q(V_1)$ and .38 for $Q(V_2)$. Suppose there are two messages, 1 and 2. The random ensemble of codes then consists of four codes.

code 1 $\Pr(\text{code 1}) = .6^2 = .36$

coding

$1 \rightarrow U_1$

$2 \rightarrow U_1$

decoding

$V_1 \rightarrow 1$

$V_2 \rightarrow 1$

code 2 $\Pr(\text{code 2}) = .6 \times .4 = .24$

$1 \rightarrow U_1$

$2 \rightarrow U_2$

$V_1 \rightarrow 1$

$V_2 \rightarrow 2$

code 3 $\Pr(\text{code 3}) = .6 \times .4 = .24$

$$1 \rightarrow U_2$$

$$V_1 \rightarrow 2$$

$$2 \rightarrow U_1$$

$$V_2 \rightarrow 1$$

code 4 $\Pr(\text{code 4}) = .4^2 = .16$

$$1 \rightarrow U_2$$

$$V_1 \rightarrow 1$$

$$2 \rightarrow U_2$$

$$V_2 \rightarrow 1$$

The distribution of mutual information under the two measures \Pr_1 and \Pr_2 is given by the following table:

	\Pr_1	\Pr_2	$I(\text{bits})$
$U_1 V_1$	$.6 \times .7 = .42$	$.6 \times .62 = .372$	$\log_2 \left(\frac{.7}{.62} \right) = .177$
$U_1 V_2$.18	.228	-.340
$U_2 V_1$.20	.31	-.308
$U_2 V_2$.20	.19	.397

The functions p_1 and $1 - p_2$, together with the sum $p_1 + (M-1)(1-p_2) = p_1 + 1 - p_2$ (since $M = 2$) are shown in Fig. 2.

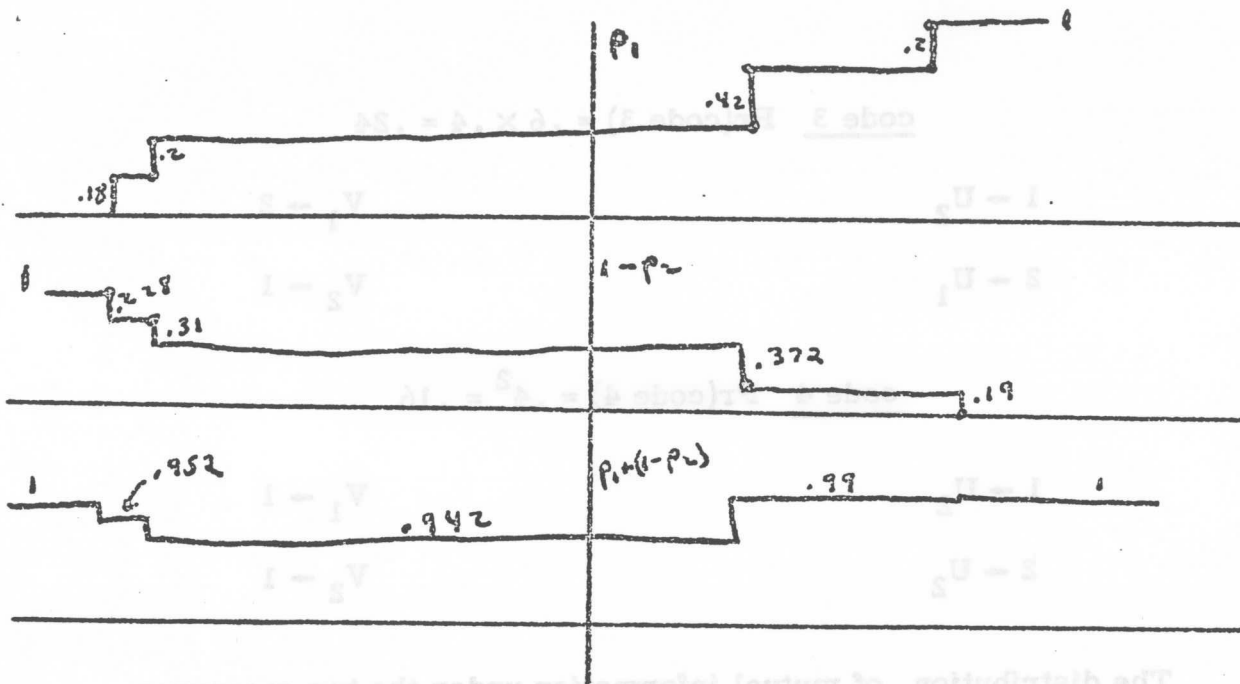


FIG. 2

This example, of course, uses a ridiculously small M and small number of input and output words in order to keep the number of codes and other complexities down. According to the theorem, the error probability will not exceed the curve $p_1(x) + 1 - p_2(x)$ at any point. The best choice of x is clearly one between $-.34$ and $.177$ for which the sum curve is $.942$. Thus we may assert that for the random ensemble $P_e \leq .942$. Actually, if the messages are equally likely, the error probability is given exactly by

$$P_e = .36(.5) + .24(.4) + .24(.4) + .16(.5) \\ = .452$$

An optimal code for two messages into this channel clearly maps them

into the two input words and gives an error probability with optimal decoding of .4. The bound of the theorem does not become very useful or significant until the number of messages and possible input words is reasonably large.

We now wish to express the bound of the lemma in terms of the assigned probabilities P_i and the transition probabilities $p_i(j)$. As noted above, I is the sum of n independent random variables (the mutual informations between corresponding letters of u and v). This is true both for the Pr_1 and Pr_2 probability measures. Thus each term of our bound relates to the problem of estimating the tail of a distribution which is the sum of n identically distributed random variables. We may conveniently estimate such tails by the Chernoff bound involving the logarithm of the moment generating function, say $\mu(s)$, of the individual random variables. Chernoff's bound states that if X_n is the sum of n such random variables, then

$$\Pr[X_n \leq n\mu'(s)] \leq e^{n(\mu(s) - s\mu'(s))} \quad s \leq 0$$

$$\Pr[X_n \geq n\mu'(s)] \leq e^{n(\mu(s) - s\mu'(s))} \quad s \geq 0$$

In our case the log moment generating function for Pr_1 , which will be called $\mu_1(s)$, is given by

$$\mu_1(s) = \log \sum_{i,j} P_i p_i(j) e^{s \log \frac{p_i(j)}{Q_j}}$$

$$= \log \sum_{i,j} P_i \frac{p_i(j)^{1+s}}{Q_j^s}$$

$$= \log \sum_{i,j} \frac{p(i,j)^{1+s}}{P_i^s Q_j^s}$$

With regard to the Pr_2 measure and estimation of $\text{Pr}_2[I \geq a]$, it is still possible to use the Chernoff bound for $s > 0$, even though I has "positive probability of being at $-\infty$." To see this, note that the moment generating function $v_2(s)$ for the Pr_2 measure is a well-defined function for $s > 0$, namely,

$$v_2(s) = \sum_{i,j} P_i Q_j e^{s \log \frac{p_i(j)}{Q_j}}$$

$$= \sum_{i,j} P_i \frac{p_i(j)^s}{Q_j^{s-1}} \quad s > 0$$

Furthermore, the generalized Chebycheff inequality with the monotone increasing function e^{sI} still holds for positive s .

$$e^{sa} \text{Pr}[I \geq a] \leq E[e^{sI}] = v_2(s)$$

$$\text{Pr}[I \geq a] \leq \frac{v_2(s)}{e^{sa}} \quad s > 0$$

In particular, setting $\alpha = \mu_2^1(s)$ where $\mu_2(s) = \log v_2(s)$, (this is the best choice to give a good bound), we obtain

$$\Pr[I \geq \mu_2^1(s)] \leq \frac{e^{\mu_2(s)}}{e^{s\mu_2^1(s)}} = e^{\mu_2(s) - s\mu_2^1(s)}$$

Note also that

$$\begin{aligned} \mu_2(s) &= \log \sum_{i,j} P_i \frac{p_i(j)^s}{Q_j^{s-1}} \\ &= \mu_1(s-1) \quad 0 < s \leq 1 \end{aligned}$$

Thus the two generating functions have, in the common range of their validity, a very simple functional relationship. It follows that $\mu_2^1(s) = \mu_1^1(s-1)$. If we wish, in using the Chernoff bound, to place the cut-off point for the tail, that is, α , at the same point, we must use s_1 and s_2 for μ_1 and μ_2 related by

$$\alpha = n\mu_1^1(s_1) = n\mu_2^1(s_2)$$

This is achieved by making $s_2 = s_1 + 1$, since then $\mu_2^1(s_2) = \mu_2^1(s_1 + 1) = \mu_1^1(s_1 + 1 - 1) = \mu_1^1(s_1)$. This is a unique solution, if we except the rather degenerate case where I is constant, since it is easily shown that in all other cases $\mu_1^1(s)$ is positive. Using s_2 and s_1 related in this manner in the Chernoff bounds, we have

$$\Pr_1[I \leq n\mu_1'(s_1)] \leq e^{n(\mu_1(s) - s_1\mu_1'(s_1))} \quad s_1 \leq 0$$

$$\begin{aligned} \Pr_2[I \geq n\mu_1'(s)] &\leq e^{n(\mu_2(s_1+1) - (s_1+1)\mu_2'(s_1+1))} \\ &= e^{n(\mu_1(s_1) - (s_1+1)\mu_1'(s_1))} \quad s_2 = s_1 + 1 > 0 \end{aligned}$$

Thus both bounds are now expressed in terms of μ_1 and its derivative with one parameter, s_1 , which must be in the range $-1 < s_1 \leq 0$. Our error probability P_e is now bounded by (writing $e^{nR} = M$ and s for s_1 , since we no longer need the subscript)

$$P_e \leq e^{n(\mu_1(s) - s\mu_1'(s))} + e^{nR} e^{n(\mu_1(s) - (s+1)\mu_1'(s))} \quad -1 < s \leq 0$$

This bound holds for any s in the allowed range. We wish to choose s to roughly minimize the bound. This is done conveniently by equating the exponents for the two terms, since the first is monotone increasing in the range (its derivative is $-ns\mu_1''(s)$) while the other is monotone decreasing (its derivative is $-n(s+1)\mu_1''(s)$). Thus we set

$$\mu_1(s) - s\mu_1'(s) = R + \mu_1(s) - (s+1)\mu_1'(s)$$

$$R = \mu_1'(s)$$

With s chosen to satisfy this, the two terms are equal and therefore

P_e reduces to twice the first term. Thus

$$\text{if } R = \mu_1^i(s) \quad -1 < s \leq 0$$

$$P_e \leq 2e^{n(\mu_1(s) - s\mu_1^i(s))}$$

If $\mu(-1) < R \leq \mu_1^i(0) = \sum_{i,j} P_i p_i(j) \log \frac{p_i(j)}{Q_j}$, there will exist a unique s in the allowed range satisfying $R = \mu_1^i(s)$. This may be seen by noting

that $\mu_1^i(s)$ is a continuous monotone increasing function of s as s ranges from -1 to 0 . Furthermore, if there are no $p_i(j) = 0$, $\mu_1^i(-1) \leq 0$.

This follows from the convexity property of the logarithm,

$$\left(\sum p_i \log x_i \leq \log \sum p_i x_i \text{ for } \sum p_i = 1 \right); \text{ we have } \mu_1^i(-1) \leq \log \sum_{i,j} P_i Q_j \frac{p_i(j)}{Q_j} =$$

$\log \sum_{i,j} P_i p_i(j) = \log 1$. Hence, in this case, for each R from 0 to the mean mutual information related to the probability assignment P_i there is a unique s .

If there are some (i,j) pairs with $p_i(j) = 0$, it is possible to have $\mu^i(s)$ approach $R_0 > 0$ as s approaches -1 . This happens, for example, in the channel Fig. 3,

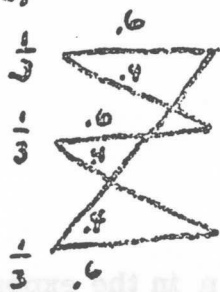


FIG. 3

for which $R_0 = \log \frac{4}{1/3} = \log 1.2 > 0$. In such a case, the bound as written above applies only between R_0 and $\mu'(0)$, the average mutual information with the given probability assignment. We may, however, extend the bound to lower rates by an argument similar to that in (1). For rates R satisfying $0 \leq R \leq R_0$ choose for the α in the lemma a value less than nI_{\min} , where I_{\min} is the smallest $\log \frac{P_i(j)}{Q_j}$ (among (i, j) pairs for which $p_i(j) > 0$, that is, not including the " $-\infty$ " cases). We then have that $\Pr_1[I \leq \alpha] = 0$, since all cases with non-zero probability in this probability measure give I values at least nI_{\min} . Also, $\Pr_2[I \geq \alpha] \leq \left(\sum_{(i, j) \in S_F} P_i Q_j \right)^n$ where S_F is the set of (i, j) pairs for which $p_i(j) \neq 0$ and hence the mutual information is finite. In fact, $\left(\sum_{(i, j) \in S_F} P_i Q_j \right)^n$ is the probability that all corresponding letter pairs of u and v have finite mutual information. If any pair fails, the u could not have been the true cause of v_1 since that letter would have involved a transition of zero probability. It follows that

$$P_e \leq e^{nR} \left(\sum_{S_F} P_i Q_j \right)^n$$

$$= e^{n \left(R + \log \sum_{S_F} P_i Q_j \right)}$$

Thus, in this region, the coefficient of n in the exponent is a linear function of the rate R of unit slope and intercept $\log \sum_{S_F} P_i Q_j$. It is

readily seen that this straight line is tangent to the curve of the previous bound at the value $R = R_0$. However the coefficient in \bar{P}_e has improved from 2 to 1.

Of course, in the ensemble of codes there must exist particular codes satisfying these same inequalities for error probability, since there is always one member of an ensemble at least as good as the average. Furthermore, if one were to choose samples from the ensemble of codes with their corresponding probabilities, then with probability at least $1 - \frac{1}{k}$ a sample will have an error probability less than or equal to $k \bar{P}_e$ for any $k > 0$. For example, with probability at least .9, the sample would have an error probability less than or equal to $10 \bar{P}_e$. This is because P_e is non-negative for each code and if the probability of exceeding $k \bar{P}_e$ were more than $\frac{1}{k}$ the average would exceed \bar{P}_e , a contradiction.

Thus a code could be generated by a Monte Carlo process or by use of a book of random numbers with high probability of not exceeding the error probability bounds excessively.

Uniformly good codes

The bounds above refer, of course, to average error probability over the different messages when all messages are used with equal probability. It is also of interest to consider uniformly good codes for which each message has a low error probability. From a code of the first type it is possible to construct a uniformly good code with slightly lower rate and poorer error probability. () In fact, if we have a code with M_1 messages and error probability less than or equal to P_{e1} (the messages used with equal probability), then at least half of the messages must have individual error probabilities less than or equal to $2P_{e1}$. This is the same combinatorial principle as used above. Thus we can find a code with $\frac{M}{2}$ (or $\frac{M-1}{2}$ if M is odd) messages and a uniform error probability bound of $2P_{e1}$. This corresponds to a rate of essentially $R - \frac{1}{n} \log 2$ and a reliability of $E - \frac{1}{n} \log 2$, where R and E are those for the given code. In other words, the same R and E curves apply if displaced in both coordinates by $\frac{1}{n} \log 2$, a quantity which rapidly approaches zero as the code length n increases.

Such uniformly good codes have the desirable feature of preserving the same bound on error probability even if the prior probabilities of different messages are changed. Indeed, they may be used if such message probabilities were entirely unknown or felt to be meaningless or non-existent in a particular situation.

Best bounds under variation of the P_i

The above bounds were deduced on the basis of an arbitrary assignment P_i of input letter probabilities. To obtain the strongest results from these bounds one may, for any particular R , vary the P_i and attempt to find the set which gives the minimum bound on error probability. Another way of looking at this is that the $E(R)$ bounding curves are found for all possible assignments and the envelope of these is used. It may be readily shown that if the channel has the "uniform input" property, then the best assignment is for all input letters to have equal assigned probability. A channel has the uniform input property if the output letters can be partitioned into a number of subsets S_1, S_2, \dots , such that each output letter in any subset S_i has the same set of transition probabilities coming into it and each input letter has the same set of transition probabilities going into S_i . A simple example is the erasure channel if both letters have the same probability of being erased.

Behavior near channel capacity

The first-order behavior of E , the coefficient of n in the error probability exponent, for rates near channel capacity may be found by a power series expansion of $E(s)$ and $R(s)$ about the point $s = 0$. Thus

$$E(s) = E(0) + sE'(0) + \frac{s^2}{2}E''(0) + \dots$$

$$= 0 + s \left(s \mu''(s) \right)_{s=0} + \frac{s^2}{2} \left(\mu''(s) + s \mu'''(s) \right)_{s=0} + \dots$$

$$= \frac{s^2}{2} \mu''(0)$$

$$R(s) = R(0) + sR'(0) + \dots$$

$$= C + s \mu'(0)$$

Eliminating s between these two relations we obtain

$$(R-C)^2 = s^2 (\mu''(0))^2$$

$$= 2 E \mu''(0)$$

$$E = \frac{(R-C)^2}{2 \mu''(0)}$$

Thus, near channel capacity, the ER curve is approximately parabolic with second derivative at C equal to $\frac{1}{\mu''(0)}$. It is readily shown that $\mu''(0)$ is the variance of mutual information, and this approximation is

related to a central limit theorem normal approximation to the distribution of mutual information near its mean. The approximate bound here near channel capacity is the same as that in (1), the two curves "osculating" at channel capacity and diverging appreciably only at lower rates.

- (1) C. E. Shannon, "Certain Results in Coding Theory for Noisy Channels"
Information and Control, Vol. 1, No. 1

RELIABLE MACHINES FROM UNRELIABLE COMPONENTS

C. E. Shannon

These notes, taken by W. W. Peterson, cover the first five lectures in the Seminar on Information Theory offered by C. E. Shannon at M. I. T., Spring term 1956. The subject matter is principally VonNeuman's "Probability Logics".

March, 1956

RELIABLE MACHINES FROM UNRELIABLE COMPONENTS

Bibliography:

VonNeuman, Probabilistic Logics, Notes by R. S. Pierce on lectures given at C.I.T., 1952. (To appear in Shannon and McCarthy, Automata Studies, Princeton University Press, 1956.)

Tsien, Engineering Cybernetics, McGraw, Hill, 1954.
The last chapter of this book is a condensed version of part of VonNeuman's paper.

Moskowitz and McLean, Some Reliability Aspects of System Design, Rome Air Development Center Report. RADC TN-55-4.

Shannon and Moore, Reliable Circuits Using Less Reliable Relays, Unpublished Bell Laboratories Report. (To appear in Journal of Franklin Institute sometime in 1956.)

1.1 Introduction

Even "near Perfect" elements may not be adequate for extremely complicated machines, or for machines whose failure might result in a catastrophe. Consider a complex machine made up of 10^6 components each of which with a probability of failure of 10^{-6} in any minute. This machine would be expected to fail once each minute, even though each particular component was expected to fail only once in ten years.

In case men's lives depend upon the successful operation of a machine, it is difficult to decide on a satisfactorily low probability of failure, and in particular, it may not be adequate to have men's fates depend upon the successful operation of single components as good as they may be.

The following methods may be used to increase reliability:

1) Complete Redesign

For example, a digital computer may be used to replace an analog computer in order to gain accuracy.

2) Improve Components

For example, most relays now have double contacts and are several orders of magnitude more reliable than single contact relays.

3) Use Error Detecting Systems

For example, numbers may be represented in a computer or data transmission system in the 2 of 5

code, in which numbers are represented by all arrangements of two ones and three zeros in five bit positions on a paper tape or other medium. A component failure would probably result in a character which had too many or too few ones and circuits which check the validity of the characters would detect most errors. Another example is the biquinary code, used in the arithmetic unit of the Bell Laboratories Relay Computer. In this code, numbers are represented by seven bits according to the following scheme:

0	01 10000	5	10 10000
1	01 01000	6	10 01000
2	01 00100	7	10 00100
3	01 00010	8	10 00010
4	01 00001	9	10 00001

In the Bell Laboratories Relay Computer, error detection was used to enable error correction by having the machine check the validity of the coded numbers after each operation and repeat any operation which resulted in erroneous results.

4. Error Correction

For example, though the individual neurons in the brain fail, the brain usually continues to operate without a serious failure for many years.

The fourth method of improving reliability is the subject of this part of the course.

As an indication of the type of analysis that will be made, consider an unreliable machine which has an input and an output which may be any one of many symbols (for example a digital computer where output is a ten digit number):

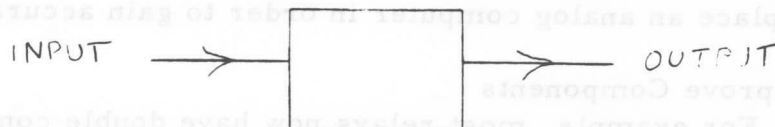


Fig. 1.1

Also consider a perfect majority device (i. e. the majority device itself never makes errors),

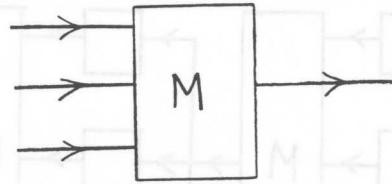


Fig. 1.2

which has three inputs. There is no output unless two of the inputs agree, in which case the common symbol is the output. Now consider three copies of the original machine with their inputs taken from the same source and their outputs connected to the majority device:

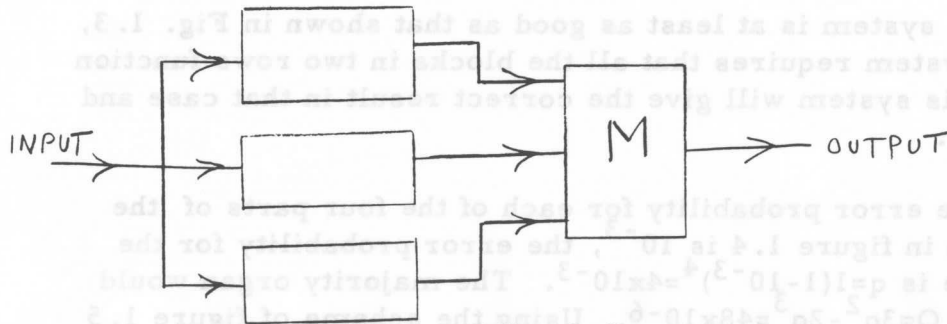


Fig. 1.3

If $p=1-q$ is the probability that the output of one of the devices is correct, if the probabilities are independent, and if the probability that two of the three erroneously agree is negligible, the probability that the device shown in figure 1.3 will give the correct output is

$$\begin{aligned} P &= p^3 + 3p^2q = (1-q)^3 + 3(1-q)^2q \\ &= 1-3q^2 + 2q^3, \text{ and} \\ Q &= 1-P = 3q^2 - 2q^3 \end{aligned} \quad (1.1)$$

If q is small, Q may be much smaller, while if q is large, Q may be considerably larger. For example, if $q=10^{-6}$, $Q=3 \times 10^{-12}$, while if $q=0.7$, $Q=0.764$. If $q=1-10^{-6}$ then $Q = 1-3 \times 10^{-12}$

Frequently a complex device is made up of many devices connected in cascade:

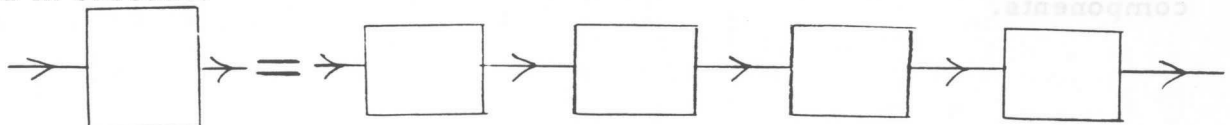


Fig. 1.4

Instead of the system considered in Figure 1.3, consider the following system:

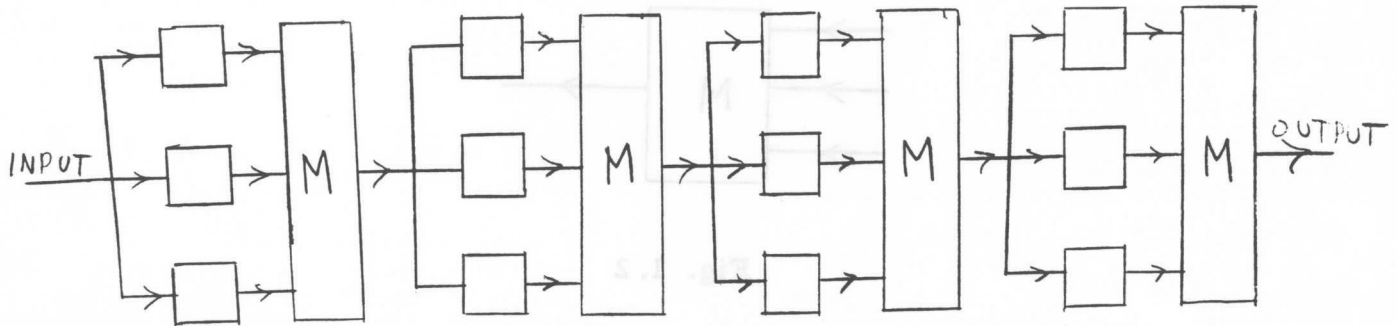


Fig. 1.5

This system is at least as good as that shown in Fig. 1.3, because that system requires that all the blocks in two rows function correctly. This system will give the correct result in that case and in many others.

If the error probability for each of the four parts of the machine shown in figure 1.4 is 10^{-3} , the error probability for the four in cascade is $q=1-(1-10^{-3})^4=4 \times 10^{-3}$. The majority organ would correct this to $Q=3q^2-2q^3=48 \times 10^{-6}$. Using the scheme of figure 1.5 for each stage $q=10^{-3}$, and hence $Q=3 \times 10^{-6}$. Four stages cascaded will give an overall probability of error $1-(1-3 \times 10^{-6})^4=12 \times 10^{-6}$, which is one fourth that obtained by the other system. If the probability for each part of the machine is taken as 0.2 instead of 10^{-3} , the resulting error probabilities for the systems shown in figures 1.3 and 1.5 are .51 and .38 respectively.

The poor features of this system compared to 1.3 are, 1) the cost of the majority devices, and 2) that in practice majority devices are not perfect, and they introduce errors also. If the machine is broken down into small enough parts the majority devices may introduce more errors than they correct.

1.2 VonNeuman's Probabilistic Logics

The basic scheme used by VonNeuman is to design the desired automaton from some sort of idealized components and then to describe a way of transforming this into a reliable automaton built from unreliable components.



The ideal components have a number of inputs and one output, as in the following diagram:

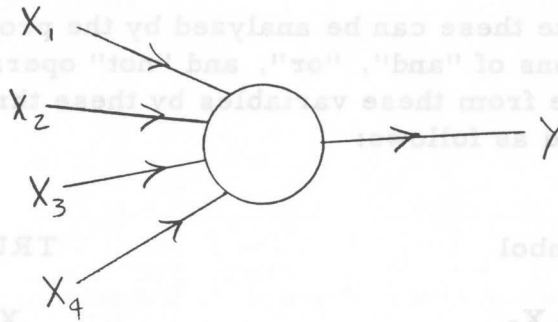


Fig. 1.6

The output variable y and all the input variables take on only the values 0 and 1. The output is a function of the input variables,

$$y = f(X_1, X_2, X_3, X_4)$$

but it is delayed by a time δ . In the following analysis all elements will be assumed to have the same delay, and this will be used as the unit of time.

In designing an automaton from these elements it is assumed that any output can be branched and connected to any number of inputs, but that two outputs are never connected together.

These ideal elements might be thought of as idealized neurons or computer logical circuits, but we will consider them simply as mathematical model without any particular interpretation.

The following special type of ideal element is particularly useful:

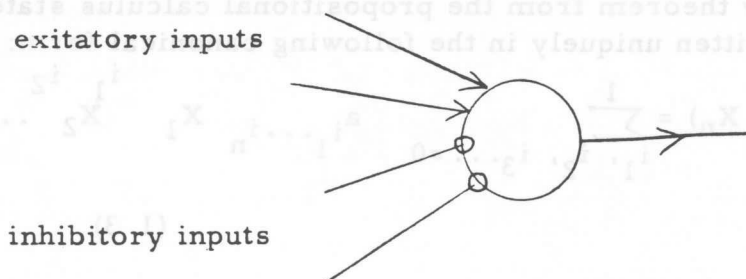


Fig. 1.7

The device may have any number of excitatory inputs and any number of inhibitory inputs. The device has a one as output only when

$$N_e - N_i \geq h,$$

where N_e is the number of excitatory inputs receiving 1's and N_i is the number of inhibitory inputs receiving 1's. A bus for supplying constant 1's and one for constant zeros will be assumed.

Devices like these can be analyzed by the propositional calculus¹, defined as combinations of "and", "or", and "not" operations on variables and polynomials made from these variables by these three operations. These operations are defined as follows:

Name Symbol TRUTH TABLE

"and" $X_1 \cdot X_2$

	X_2	0	1
X_1	0	0	0
1	1	0	1

"or" $X_1 + X_2$

	X_2	0	1
X_1	0	0	1
1	1	1	1

"not" X_1'

X_1	1	0
X_1'	0	1

One noteworthy theorem from the propositional calculus states that any polynomial can be written uniquely in the following canonical form:

$$p(X_1, X_2, \dots, X_n) = \sum_{i_1, i_2, i_3, \dots = 0}^1 a_{i_1 \dots i_n} X_1^{i_1} X_2^{i_2} \dots X_n^{i_n} \quad (1.3)$$

where $X^0 = X$ and $X^1 = X'$. The coefficients $a_{i_1 i_2 \dots i_n}$ are essentially the truth table for p , and the proof consists essentially of noting that if the truth table of a polynomial p is used as coefficients, the canonical form will agree with p .

-
1. Couturat, The Algebra of Logic, Paris, 1905
Birkoff & MacLane, A Survey of Modern Algebra, MacMillan, 1953.

Ideal elements which do the "and", "or", and "not" operations can be formed as follows:

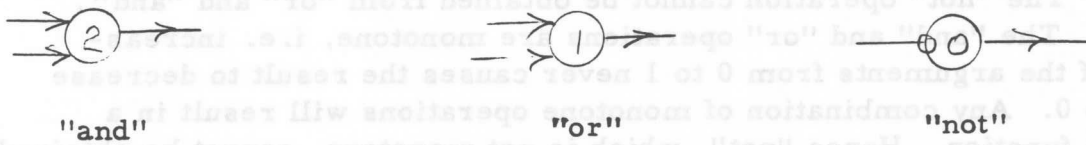


Fig. 1.8

Of these a device with arbitrary function as output can be built. This can be proved by an induction on the number of "and", "or", and "not" operations in the expression. For $n=1$, the function can be formed with one of the basic elements shown in figure 1.8. Assuming that the statement is true for all functions containing no more than n operations, the device for a function with $n+1$ operations can be constructed as follows: consider the $(n+1)^{st}$ operation. Its operand(s) certainly contain no more than n operations, and therefore, devices can be constructed which correspond to them. The outputs from these devices can be combined using the basic element corresponding to the $(n+1)^{st}$ operation to give the required device.

Any function can be generated in this manner, but there will be a delay. For any given function there is a minimum delay. The same function can be obtained with arbitrary delay greater than the same function by using any number of "and" circuits as unit delay elements.



Fig. 1.9 - Delay Element

In order to simplify the mathematical analysis, we wish to reduce the number of types of elements to a minimum. By using DeMorgan's Theorem:

$$(x+y)' = x'y', \text{ or } x+y = (x'y')'$$

the "or" can be obtained from "and" and "not" elements. Thus

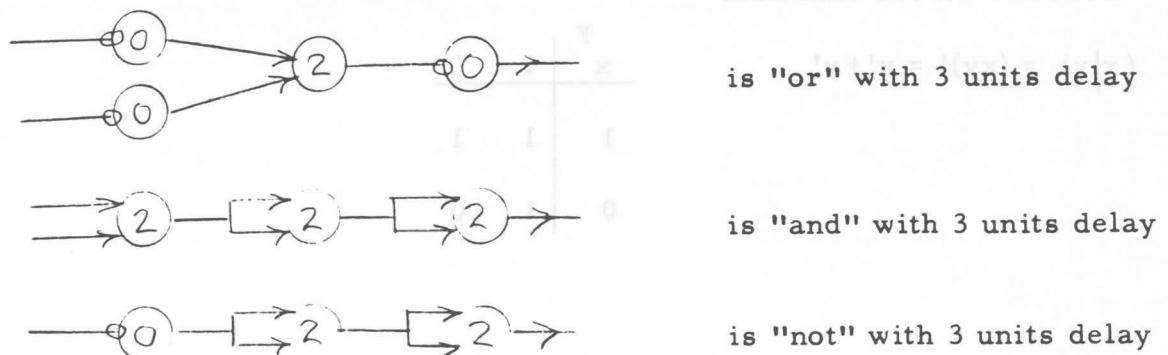


Fig. 1.10

and these could be used as basic elements. Similarly, "and" can be obtained from "or" and "not".

The "not" operation cannot be obtained from "or" and "and", however. The "and" and "or" operations are monotone, i.e. increasing one of the arguments from 0 to 1 never causes the result to decrease from 1 to 0. Any combination of monotone operations will result in a monotone function. Hence "not", which is not monotone, cannot be obtained from any combination of "and" and "or" operations.

There is another way of organizing an automaton which does make "and" and "or" sufficient for obtaining all polynomials. It is the "double line trick" in which each variable is represented by two lines. A one is represented by a 1 on the first line of the pair and an 0 on the second, while a zero is represented by the opposite. With this convention, the "and", "or", and "not" can be obtained from "and" and "or" elements as follows:

"not"

"and"

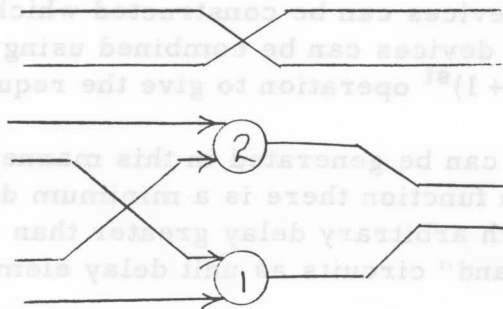


Fig. 1.11

The "or" can be obtained by using DeMorgan's Theorem, i.e., by twisting both input lines and the output line. (It turns out that this is equivalent to interchanging the basic "and" and "or" elements in the diagram of the "and" circuit.) Note that the "not" circuit needs a delay of one unit to make it have the same delay as the "and" and "or".

It was discovered by Scheffer that there are single functions from which all these, "and", "or", and "not" can be obtained. One is the Scheffer stroke function:

$$(x|y) = (xy)' = x' + y'$$

x	y	
	1	0
1	1	1
0	1	0

(1.4)

In terms of it,

$$x' = (x|x)$$

$$x \cdot y = (x|y) | (x|y)$$

$$x + y = (x|x) | (y|y) \quad (1.5)$$

The stroke function can be represented by an element of the following type:



Fig. 1.12

and "and", "or" and "not" circuits can be built from Scheffer stroke elements according to the above formulas. Note, however, that the "not" circuit requires one stroke function and hence only one unit delay, while the "and" and "or" circuits require 2 stroke functions cascaded, and hence two units of delay. This time the delay cannot be equalized. The stroke function is anti-monotone. In a device made from stroke functions; any points removed from the input by one stroke element will be anti-monotone. Any points two levels deep are monotone, etc. Thus the "not" which is anti-monotone cannot be obtained at the same level, and hence time delay, as the "and" and "or" which are monotone.

Since "and" and "or" can be obtained at the same level, we can use the double-line trick to obtain "and", "or", and "not" in terms of stroke elements.

Another element of interest is the "majority organ":



Fig. 1.13

It is monotone, but the "and" and "or" can be obtained from it as follows:

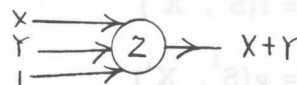
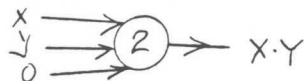


Fig. 1.14

and hence the majority organ is universal when used with the double line trick.

From the elements we have discussed we can build black boxes of the following kind, with any given set of propositional functions $f_i(x_1, \dots, x_n)$ $i=1, 2, \dots, m$,

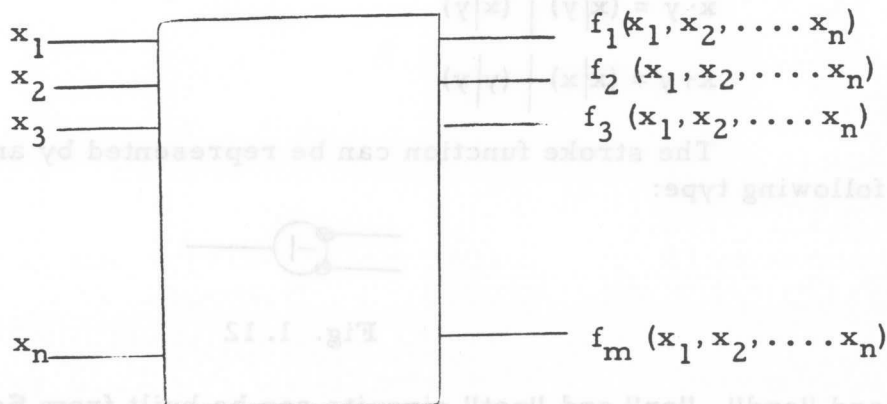


Fig. 1.15

and we can line up the outputs by using delay elements. The notation can be shortened by writing X for the vector $(x_1, x_2, x_3, \dots, x_n)$ and $F(X)$ for the vector function (f_1, f_2, \dots, f_n) .

A more general type of machine has outputs which depend not only on the input but also in some way upon the previous history of the device. One very general model of such a device is the "finite state transducer" which is a satisfactory representation of a digital computer, for example. At each time interval i it is given an input (vector) X^i , it has a state (vector) S^i which can assume a finite number of possible values, and produces an output (vector) Y^i :

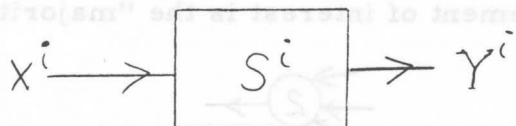


Fig. 1.16

They are related by the following equations:

$$\begin{aligned} S^{i+1} &= f(S^i, X^i) \\ Y^i &= g(S^i, X^i) \end{aligned} \quad (1.6)$$

The relationship between finite state transducers and devices made of the basic elements is made clear by the following two theorems:

THEOREM Any device made by combining a finite number of basic elements is a finite state transducer.

If the output of the j^{th} element at time i is denoted by s_j^i , then certainly s_j^i is a function of s_j^{i-1} and the input, and can be interpreted as the components state vector S^i . Then the output (vector) Y^i is certainly a function of S^i and X^i , since the outputs must come either from an element or directly from the input.

THEOREM Given the equations for a finite state transducer, such a transducer can be built of basic "and", "or" and "not" elements, (or any other set of universal components) except that the interval of time between inputs and between outputs will be some multiple of the unit delay, and the outputs Y^i may be delayed by some multiple of the unit delay.

Suppose there are k states. They can be represented by the binary numbers from zero to $k-1$, and the binary digits of these numbers can be used as the components of the state vector S^i . Then the original given equations for the transducer become equations in binary variables:

$$S^{i+1} = f(S^i, X^i)$$

$$Y^i = g(S^i, X^i)$$

Black boxes of the type shown in Fig. 1.15 can be made corresponding to each of these equations, but they will have delays. Suppose each output from the first is delayed d_1 units, and each output from the second d_2 units. Then the finite state transducer is obtained by connecting them as follows:

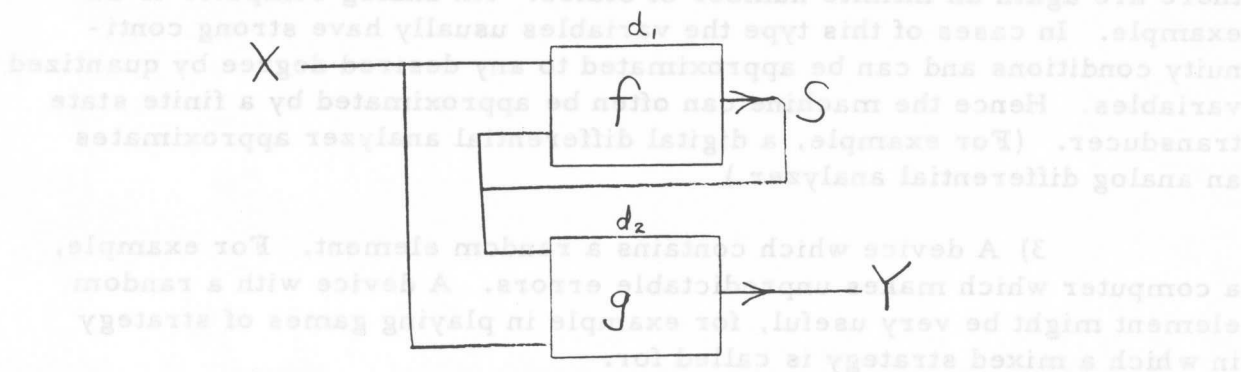


Fig. 1.17

If an input is entered at times $0, d_1, 2d_1, 3d_1$, etc., the inputs will synchronize with the state variables coming out of the f box to satisfy the first equation, and the second box will produce the outputs according to the second equation, at times d_2, d_1+d_2 , etc.

(Actually, the machine could simultaneously process d_1 input sequences starting at times $0, 1, 2, \dots, d_1 - 1$ respectively and produce the d_1 output sequences similarly meshed).

Problems:

1. a. Prove that it is possible to build a device with two inputs and one output which produces the sum of the input binary numbers for numbers of arbitrary length.
b. Prove that this is not possible for multiplication. (It is also not possible for square root)
2. Design a device from an infinite number of "and", "or", and "not" elements which is equivalent to a universal Turing machine.

The following are examples of types of machines which are not finite state transducers:

- 1) A device which has an infinite number of states, for example, a Turing machine with its infinite tape.
- 2) A device in which continuous variables occur, and hence there are again an infinite number of states. An analog computer is an example. In cases of this type the variables usually have strong continuity conditions and can be approximated to any desired degree by quantized variables. Hence the machine can often be approximated by a finite state transducer. (For example, a digital differential analyzer approximates an analog differential analyzer.)
- 3) A device which contains a random element. For example, a computer which makes unpredictable errors. A device with a random element might be very useful, for example in playing games of strategy in which a mixed strategy is called for.

Now we shall consider automata constructed of basic elements which sometimes fail, with the failures occurring according to some probability measure. We could assume a completely general probability measure on the space of all parts of the machine, i. e., we could include all sorts of correlation. Some correlation certainly occurs in real machines. Vacuum tube failures, for example, are frequently the result of the application of improper voltages. Since the voltages are usually applied to many tubes, a number of failures may result from one occurrence of improper voltage, and hence correlation appears among the failures. Likewise, in a relay machine most failures result from dust. The fact that one relay fails is an indication that dust is present and other relays are likely to fail also.

To assume a completely general probability measure would make the problem so difficult mathematically that we could hardly expect to accomplish anything.

We shall assume that the errors which occur in the different basic elements are independent. We shall use majority organs as basic elements and assume that the probability of erroneous output is ϵ regardless of the number of 1's at the input. (A physical realization of the majority organ might not have this property. It might be more reliable when the inputs are zeros than when they are ones, or it might be more reliable when all inputs are alike than with two ones and a zero or vice versa). (A possible generalization of this would be to assume that the probability of failure of any element is less than ϵ regardless of the number of 1's at the input and regardless of the state of other parts of the machine).

If the output from the machine appears on one line, the probability of error of the output is at least ϵ (except in the trivial cases in which it comes directly from the input or from a zero or one bus) simply because the output must come from a majority organ which has probability of error of ϵ .

If η_1, η_2, η_3 , are upper bounds on the error probabilities of the three inputs to a majority organ, then the probability of error for the output of the majority organ satisfies the inequality

$$\eta^* \leq \eta_1 + \eta_2 + \eta_3 + \epsilon. \quad (1.7a)$$

This gives an absolute upper bound on the error probability, regardless of any correlations which may exist. It does not offer any hope of improvement since it never promises any decrease in error probability at the output of the majority organ.

If we assume, 1) that these probabilities are independent, and, 2) that the three inputs agree if they are correct, a stronger result can be obtained. The probability that at least two of the inputs are incorrect is then

$$\begin{aligned} \Theta &= \eta_1 \eta_2 (1-\eta_3) + \eta_1 \eta_3 (1-\eta_2) + \eta_2 \eta_3 (1-\eta_1) + \eta_1 \eta_2 \eta_3 \\ &= \eta_1 \eta_2 + \eta_1 \eta_3 + \eta_2 \eta_3 - 2\eta_1 \eta_2 \eta_3. \end{aligned} \quad (1.7b)$$

The probability of an error in the output is the probability that either 1) at least two inputs are incorrect, and the majority organ works properly, or 2) at least two inputs are correct and the majority organ makes an error, (but not both,) hence,

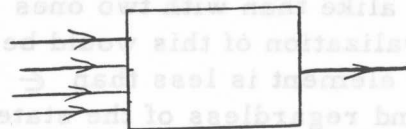
$$\eta^* = \theta(1 - \epsilon) + \epsilon(1 - \theta). \quad (1.8)$$

If $\eta_1 = \eta_2 = \eta_3 = \eta$,

$$\theta = 3\eta^2 - 2\eta^3, \text{ and}$$

$$\eta^* = \epsilon + (1 - 2\epsilon)(3\eta^2 - 2\eta^3) \quad (1.9)$$

Now consider a machine which makes errors:



Make three identical copies and connect the outputs to a majority organ.

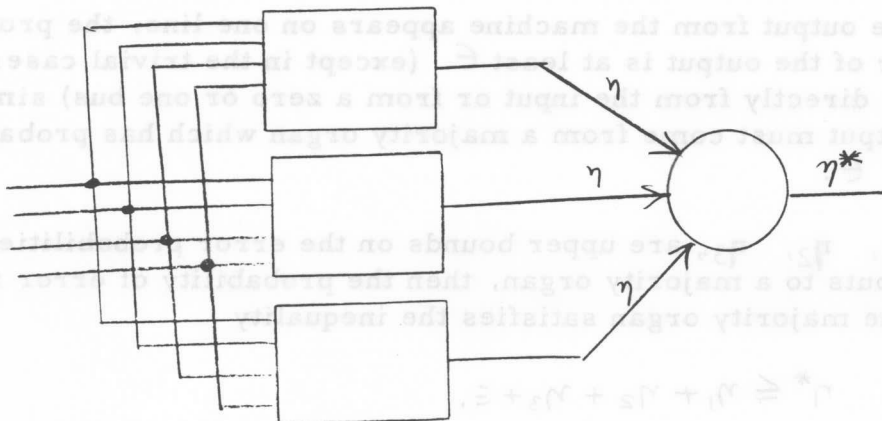


Fig. 1.18

Errors in the outputs can now be considered independent because they occur in different machines. Also, the outputs will agree if they are correct. Hence the above formula applies. But if this is to work, the output error probability η^* must be less than the input error probability η .

From equation (1.9) it can be shown that η^* considered as a function of η passes through the point $(1/2, 1/2)$. It has zero slope when $\eta = 0$ or 1 . The graph looks something like this:

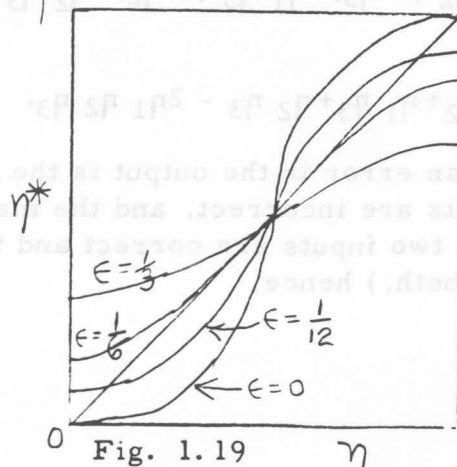


Fig. 1.19

The curve is tangent to the diagonal at the center for $\epsilon = 1/6$. In order for η^* to be less than η , the curve must lie below the diagonal for $\eta < 1/2$, and hence ϵ must be less than $1/6$. The curve for $\epsilon = 1/12$ crosses the diagonal at $\eta = 1/2$ and also at $\eta = 0.15$ and 0.85 approximately. For $\eta < 0.15$, $\eta^* > \eta$ and the error probability is increased. For $0.15 < \eta < 1/2$, on the other hand, $\eta^* < \eta$ and the error probability is decreased. In either case iterating the procedure makes the error probability approach 0.15 . Thus this crossing acts as a kind of stable point.

Now let us consider in more detail the design of a machine. Consider a machine (designed on the assumption of error free components) which has no feedback in it. It would be of the type shown in Figure 1.15.

Theorem Given an error-free design Q with no feedback we can construct an equivalent machine Q^* (with added delay). Each element of Q^* has error probability ϵ , but the whole machine has error probability less than $\eta(\epsilon)$. $\eta(\epsilon)$ is independent of machine complexity and approaches zero as ϵ approaches zero.

The proof is by an induction on the depth n of the machine. The theorem is obviously true for $n=0$, since that would mean all outputs come directly from inputs or zero or one buses. Assume that it is true for $n=k$, and consider a machine of depth $k+1$. Since there is no feedback, all the outputs from the majority organs at the greatest depth must be connected only to outputs of the machine. If these elements are removed, the rest of the machine will have depth k :

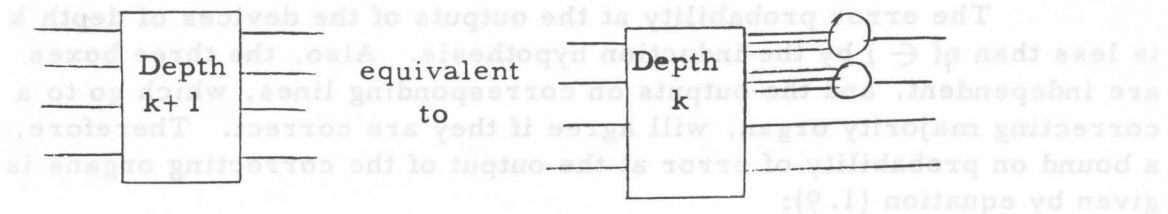


Fig. 1.20

Now by the induction hypothesis we can build this machine of depth k with error probability less than $\eta(\epsilon)$. Build three copies of it. Then connect each set of these corresponding outputs to a majority organ. Finally connect these outputs to the $k+1$ layer of majority organs:

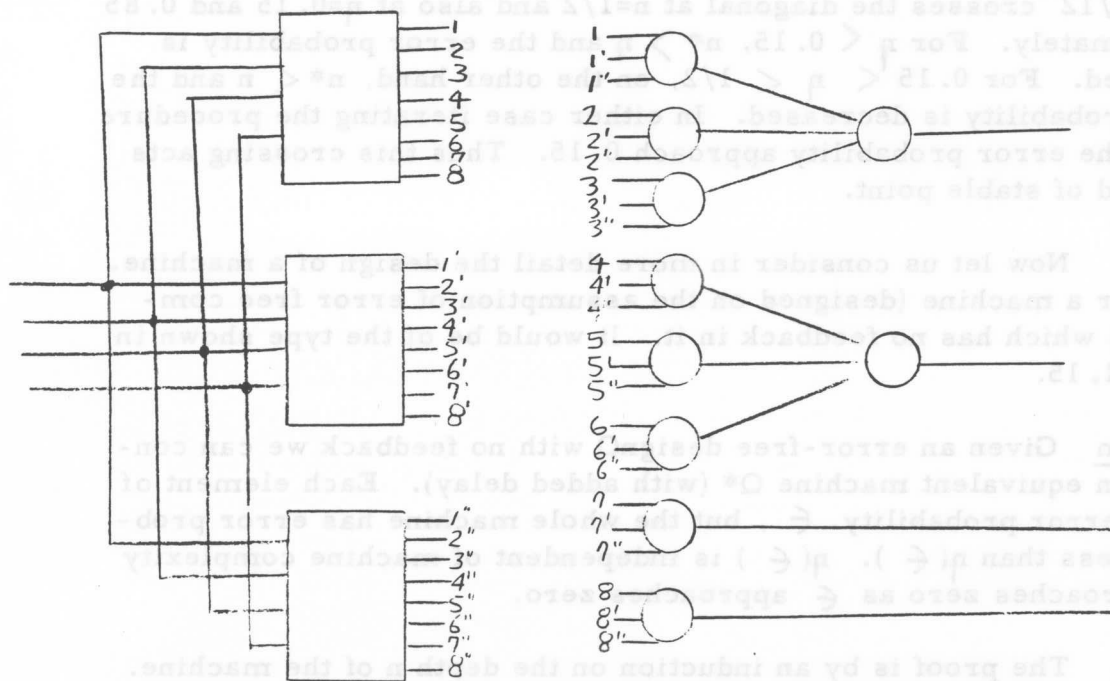


Fig. 1.21

The error probability at the outputs of the devices of depth k is less than η (\in) by the induction hypothesis. Also, the three boxes are independent, and the outputs on corresponding lines, which go to a correcting majority organ, will agree if they are correct. Therefore, a bound on probability of error at the output of the correcting organs is given by equation (1.9):

$$\eta^* = \epsilon + (1 - 2\epsilon)(3\eta^2 - 2\eta^3)$$

The probabilities at the inputs to the computing organs are less than η^* , but they are not necessarily independent, nor need they agree if correct. Therefore, we use equation (1.7a):

$$P_{\epsilon} \leq 3\eta^* + \epsilon \quad (1.10)$$

Combining these equations, we find

$$P_{\epsilon} \leq 4\epsilon + 3(1 - 2\epsilon)(3\eta^2 - 2\eta^3) \quad (1.11)$$

for the probability of error at the output of the device as shown in figure 1.21. In order to complete the proof we have to assure that $P_e \leq \eta$, and it is this requirement that defines the function $\eta(\epsilon)$.

The curve of equation (1.11) is similar to the curve of equation (1.9), except that

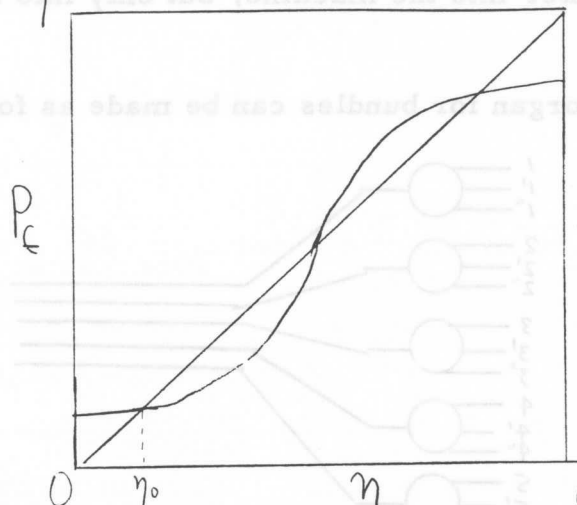


Fig. 1.22

the critical ϵ at which the curve becomes tangent to the diagonal is approximately .0073. Clearly, if β is any number such that $\eta_0 \leq \beta \leq 1/2$, (where η_0 is the point where the curve crosses the diagonal, then whenever $\eta < \beta$, P_e will be less than β also. Therefore the function $\eta(\epsilon)$ can be any function which satisfies the inequality $\eta_0 \leq \eta(\epsilon) \leq 1/2$ for all ϵ . In particular $\eta(\epsilon) = \eta_0$ is acceptable.

Note that the fact that there is no feedback plays a part in this proof.

One variation on this system would be to iterate this triplicating a number of times at each level of depth of the device. This will permit using majority organs which have error probabilities greater than .0073; it is possible to have ϵ at least as large as .125, and probably very near $1/6$.

Adding one level of depth triplicates all previous equipment and adds some, so that the redesigned machine contains much more than 3^n times the amount of equipment involved in the first level of depth. Even for modest values of n , this makes a fantastically large machine.

Now we will consider another system, which is less sensitive to errors on individual lines. It is called "multiplexing of lines". With this system, one line in the original device is represented by a "bundle" of many lines, most of which will carry a one when the corresponding line in the original machine carries a one, and most would carry a zero when

the corresponding line carries a zero. The threshold level will be denoted by δ : if the fraction of lines excited in a bundle is less than δ , the bundle will be interpreted as carrying a zero. If the fraction is greater than $1 - \delta$, it will be interpreted as a 1. If it is between δ and $1 - \delta$, the result will be considered as uncertain. This "fiduciary level" δ , does not enter into the machine, but only into the analysis of the machine.

A majority organ for bundles can be made as follows:

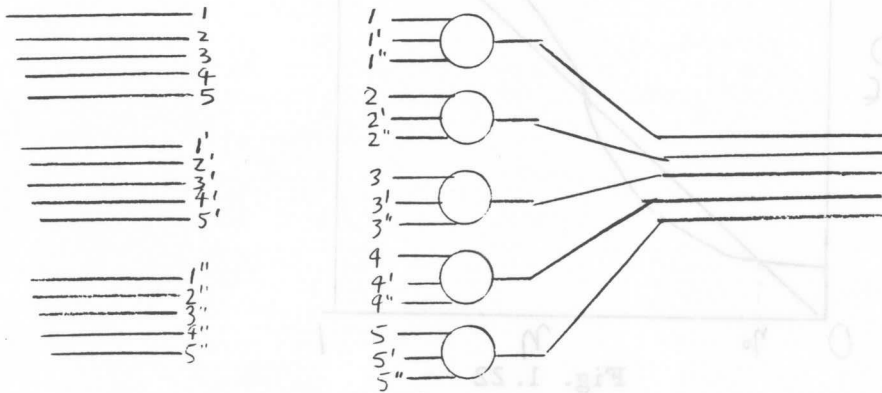


Fig. 1.23

If all the lines in each of two bundles are excited, then except for majority function errors, all outputs will be excited. Similarly, it works for all zeros on two bundles, so that the device works roughly as it should.

Now suppose fractions a , b , and c respectively of the three inputs are in error. Neglect errors in the majority organs. Also suppose that the first two bundles carry 1's, while the third carries a zero. The largest number of errors in the output would be achieved by having all the zeros in the first bundle matched with ones of the second and zeros of the third. Similarly all the zeros of the second bundle should be matched with ones of the first and zeros of the third. This would make a fraction $a+b$ of the outputs wrong. The same would apply if the first two were zeros and the third a one, by the duality between zero and one in the majority organ.

If all three inputs are ones, then there will be the most errors in the output if every error in the output is caused by two erroneous input lines. The number of errors in the output bundle certainly cannot exceed half the total number of erroneous lines in all three bundles at the input. Thus the fraction d of errors in the output is

$$d \leq 1/2 (a+b+c).$$

(1.12)

(This can almost be achieved if a , b , and c are the sides of a triangle. Otherwise, d is less than the sum of the smallest two of a, b, c .)

If $a=b=c$, the bound on errors at the output is $2a$ for the first case (not all three inputs the same) and $3/2a$ for the second (all three inputs the same). The bound we have on error probability at the output of the organ ($2a$ or $3/2a$) is thus greater than the bound a on error probability at the input.

The error probability might decrease if we consider an average situation instead of the worst possible situation. Consider the case in which all three bundles are carrying the same symbol (0 or 1), and take the average over all permutations of all the erroneous lines in each of the input bundles. Then the probability of at least two erroneous inputs to any given majority element is

$$\begin{aligned} d &= ab(1-c) + ac(1-b) + bc(1-a) + abc \\ &= ab + bc + ca - 2abc \end{aligned} \quad (1.13)$$

and this will also be the mean fraction of lines excited in the output. (Assuming $\epsilon = 0$). In any particular case some variation from this would be expected.

$$\begin{aligned} \text{If } a=b=c, \\ d &= 3a^2 - 2a^3, \end{aligned} \quad (1.14)$$

the same equation which occurred before (Equations (1.1) and (1.9) with $\epsilon = 0$), but for a different reason.

VonNeuman proposed the following as a system for restoring the level of the fraction of lines excited in a bundle (to mean 0 or 1). Each line in a bundle is to be split three ways, to get $3n$ lines. Then these would be put through a "random permutation" black box. The outputs would be connected to majority organs:

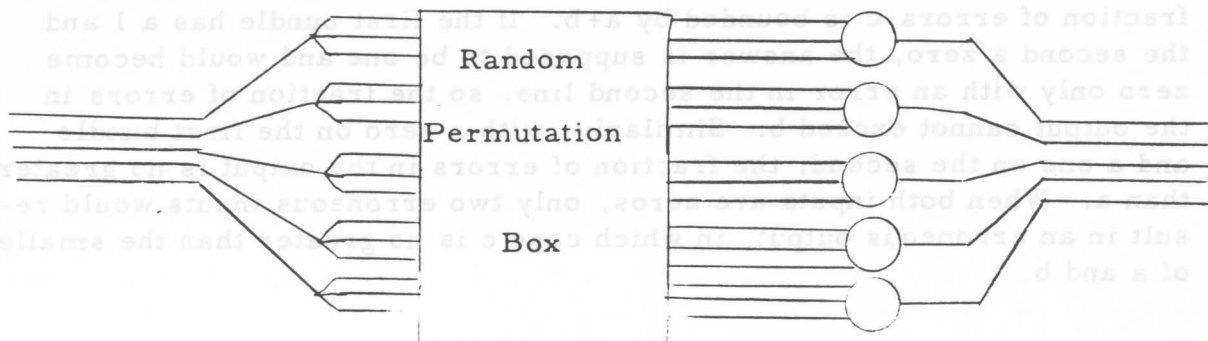


Fig. 1.24

This black box might be wired so that each input is connected to one and only one output according to a table of random numbers. The idea is to achieve the effect of independence of the inputs to any one majority organ so that formula (1.13) applies. There is no rigorous proof that this can be done, but it seems very plausible.

The same analysis can be done with Scheffer stroke organs. It could be done indirectly by noting that a majority organ can be constructed from any set of universal organs, and hence all results which hold for majority organs hold also for any other set of universal organs. The error probability ϵ would have to be that for the constructed majority organ, of course, rather than that for the basic elements themselves. The analysis for the stroke organs is simple and interesting enough to do in detail.

The stroke function for a bundle can be constructed as follows:

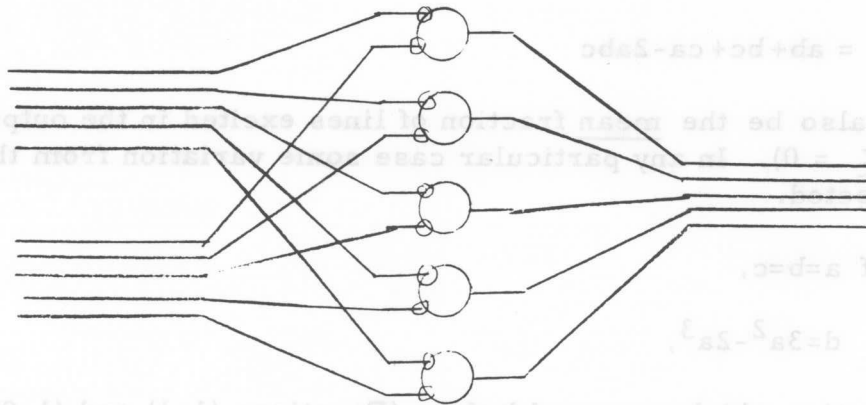


Fig. 1.25

If both inputs are supposed to be on, the result should be zero and an error will occur if either input to an organ is off. Therefore the number of errors in the output can be as great as the sum of the number of errors on both inputs, but it cannot exceed this number. Therefore the fraction of errors c is bounded by $a+b$. If the first bundle has a 1 and the second a zero, the answer is supposed to be one and would become zero only with an error in the second line, so the fraction of errors in the output cannot exceed b . Similarly, with a zero on the first bundle and a one on the second, the fraction of errors in the output is no greater than a . When both inputs are zeros, only two erroneous inputs would result in an erroneous output, in which case c is no greater than the smaller of a and b .

If the fraction of inputs excited is a for both inputs and the average over all permutations is considered, then an output will be 0 only if both inputs are 1 to a particular stroke organ, and this would occur on the average for a fraction a^2 of the line. Therefore the fraction of lines excited at the output would be

$$c = 1 - a^2 \quad (1.15)$$

The curve looks like this:

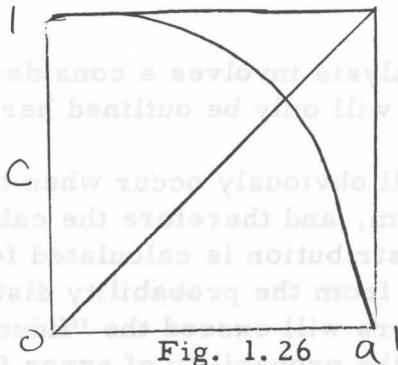


Fig. 1.26

It does not restore, but rather reverses. To get restoring, the process should be done twice. The effect of the iteration can be found by substituting (1.15) in itself as the argument, i.e.

$$a^* = 1 - (1 - a^2)^2 = 2a^2 - a^4 \quad (1.16)$$

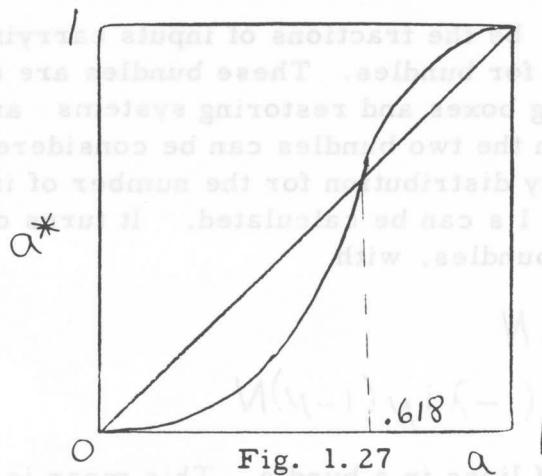


Fig. 1.27

To review the design procedure, we start with a single line machine designed for error free stroke elements. Each line is replaced by a bundle. Each organ is replaced by a bundle organ, followed by a pair of cascaded Scheffer stroke restoring organs with "random permutation" black boxes.

Until now we have not considered errors in the basic organs. Furthermore we have not considered the effect of dispersion in the number of lines excited in a bundle. We have shown only that the average number of lines excited in a bundle can be kept under control. The probability that the deviation from this average value will cause failure must be considered. The number of lines excited has a distribution similar to a binomial distribution, and in our case as with the binomial distribution, the dispersion can be made very small by making the number of lines per bundle very large.

The rest of the analysis involves a considerable amount of algebraic manipulation, and it will only be outlined here.

The worst case will obviously occur when the fraction of errors on each input line is a maximum, and therefore the calculation is made for that case. The probability distribution is calculated for each of the combinations of input signals, and from the probability distributions, the probability that the fraction of errors will exceed the "fiduciary level" can be found. This will be called the probability of error for this part of the machine. This is really a conservative estimate of probability of error, since the machine might, and probably would, function perfectly well and have the fraction of errors in the outputs less than the fiduciary level even though the fiduciary level might be exceeded at certain points within the machine.

Let λ and ν be the fractions of inputs carrying 1's on the two inputs of a stroke function for bundles. These bundles are assumed to come from different randomizing boxes and restoring systems and hence the arrangement of the errors in the two bundles can be considered random and independent. The probability distribution for the number of individual stroke elements "excited" by two 1's can be calculated. It turns out to be approximately normal for large bundles, with

$$\begin{aligned} \text{mean} &= \lambda \nu N \\ \text{variance} &= \lambda(1-\lambda)\nu(1-\nu)N \end{aligned} \tag{1.17}$$

where N is the number of lines in a bundle. This mean is consistent with the average fraction of lines excited at the output calculated before, as given in equation (1.15).

Assuming that the individual stroke elements have probability ϵ of an error, the probability distribution of errors in the output bundle is still approximately normal for large bundles, with

$$\begin{aligned} \text{mean} &= \left[\lambda \mu - 2\epsilon \left(\lambda \mu - \frac{1}{2} \right) \right] N \\ \text{variance} &= \left[(1-2\epsilon)^2 \lambda(1-\lambda) \mu(1-\mu) + \epsilon(1-\epsilon) \right] N \end{aligned} \quad (1.18)$$

Now in order to continue the calculation, a fiduciary level must be chosen. It is clear that it should not be either too near zero or too near $1/2$, and therefore there must be some optimum choice of δ . VonNeuman chose $\delta = 0.07$, but he does not explain how he arrived at that value. For this value of δ , ϵ must be no larger than $.0107$. If $\epsilon > .0107$, it is not possible to make the probability of error approach zero by making N approach infinity. If ϵ is near $.0107$ the approach to zero as a function of N would be slow. Von Neuman chose to make calculations for $\epsilon = .005$.

Substituting these values in the probability distribution and integrating over the tail gives the following approximate expression for probability of error.

$$P_E \doteq \frac{6.4}{\sqrt{N}} 10^{-\frac{8.6N}{10,000}} \quad (1.19)$$

It is convenient to define a quantity E called reliability, as

$$E = -\log P_E. \quad (1.20)$$

For this situation,

$$E \doteq \frac{8.6N}{10000} - \log 6.4 + \frac{1}{2} \log N \doteq \frac{N}{1000} \quad (1.21)$$

if the base 10 is chosen.

For a human brain, an overestimate of the mean free number of single operations between errors is

$$\begin{aligned} & (10^{13} \text{ components}) \cdot (10 \text{ operations per second}) \\ & \cdot (3600 \cdot 24 \cdot 365 \text{ seconds per year}) \cdot (100 \text{ years}) \\ & \doteq 10^{23} \end{aligned}$$

This would require E to be 23, or N to be roughly 23,000. Now a correction is necessary; since we estimated that the "brain" had 10^{13} elements and since it is already multiplexed according to our assumption, the error free design must have had roughly $10^{13} \times 23,000 = 5 \times 10^9$. Using this figure in place of 10^{13} , we find N to be roughly 18,500, a modest change in N .

Each prototype element requires $3N$ elements, N for calculating, and $2N$ for restoring. Thus the number of elements in the machine is multiplied by something like 55,000.

Cutting these figures to a minimum, perhaps the number of elements in the brain should be taken as 10^9 and the time to a matter of days instead of 100 years gives an estimate of the number of operations between errors of roughly 10^{14} instead of 10^{23} . This doesn't even cut N to one half its previous value.

Of course the forgoing statements aren't meant to imply that the brain is organized along these lines. In fact it almost certainly is not. For small values of N this multiplexing makes the error probability greater, and therefore gradual evolution of a system like this would be unlikely to occur.

As another example, consider a computing machine of say 1000 elements with 10^5 operations per second and perhaps a requirement of 3 hours mean free time between errors. The mean number of operations between errors would be roughly

$$1000 \times 10^5 \times (3600 \times 3) = 10^{12},$$

so that E should equal about 12. This would require an N of about 12000, or 36,000 times as many elements as in the original design. Of course this assumed $\epsilon = .005$, which is very poor compared to actual computer elements.

The Portfolio Problem

[106]

The following analysis, due to John Kelly, was inspired by news reports of betting on whether or not the contestant on the TV program "64,000 Questions" would win. It seems that one enterprising gambler on the west coast, where the program broadcast is delayed three hours, was receiving tips by telephone before the local telecast took place. The question arose as to how well the gambler could do if the communication channel over which he received the tips was noisy.

Consider first the case where there are two equally likely events on which the gambler may bet with 1-1 odds. Suppose the gambler receives tips which he knows are correct. Then he can double his money each time he bets. If he bets V_0 dollars, after n bets he will have $V_n = V_0 2^n$ dollars. This suggests the definition of effective interest rate r :

The Portfolio Problem and How to Pay the Forecaster

These notes, taken by W.W. Peterson, cover several

lectures in the Seminar on Information Theory offered

by C.E. Shannon at M.I.T., Spring Term, 1956.

An alternative approach would be for the gambler to bet a fraction e of his money on each bet. If he starts with V_0 and wins on the first bet he will have $2eV_0 + (1-e)V_0 = (1+e)V_0$. If he loses he will have only $(1-e)V_0$. It is clear that each successive win multiplies his holdings by $1+e$ while each successive loss multiplies his holdings by $1-e$. After W wins and L losses he will have

$$V_n = (1+e)^W (1-e)^L V_0$$

dollars. The effective interest rate is

$$r_n = \frac{W}{n} \log(1+e) + \frac{L}{n} \log(1-e)$$

1. With interest rate r , after n periods,

$$V_n = V_0 (1+r)^n$$

Substituting this in (1) above gives

$$r = \frac{1}{n} \log_2 (1+e)^W (1-e)^L = \log_2 (1+e)^{\frac{W}{n}} (1-e)^{\frac{L}{n}}$$

Thus r is a simple monotone function of the interest rate in the ordinary sense, and maximizing r is equivalent to maximizing r .

The Portfolio Problem

The following analysis, due to John Kelly, was inspired by news reports of betting on whether or not the contestant on the TV program "\$64,000 Question" would win. It seems that one enterprising gambler on the west coast, where the program broadcast is delayed three hours, was receiving tips by telephone before the local telecast took place. The question arose as to how well the gambler could do if the communication channel over which he received the tips was noisy.

Consider first the case where there are two equally likely events on which the gambler may bet with 1-1 odds. Suppose the gambler receives tips which he knows are correct. Then he can double his money each time he bets. If he starts with V_0 dollars, after n bets he will have $V_n = V_0 2^n$ dollars. This is equivalent to an interest rate of 100%. This suggests the definition of effective interest rate r :

$$r = \frac{1}{n} \log_2 \frac{V_n}{V_0}. \quad (1)$$

Now consider the case in which the tips have only probability $p > 1/2$ being correct. Probability theory states that the expected winnings are greatest when the gambler always bets all his money on the event which his tip indicates is most likely to occur. His probability of going broke after n bets, however, is equal to $(1-p)^n$, and this approaches ~~zero~~ _{0 as ϵ} as n approaches infinity.

An alternative approach would be for the gambler to bet a fraction e of his money on each bet. If he starts with V_0 and wins on the first bet he will have $2eV_0 + (1-e)V_0 = (1+e)V_0$. If he loses he will have only $(1-e)V_0$. It is clear that each successive win multiplies his holdings by $1+e$ while each successive loss multiplies his holdings by $1-e$. After W wins and L losses he will have

$$V_n = (1+e)^W (1-e)^L V_0 \quad (2)$$

dollars. The effective interest rate is

$$r_n = \frac{W}{n} \log (1+e) + \frac{L}{n} \log (1-e) \quad (3)$$

1. With interest rate i , after n periods,

$$V_n = V_0 (1+i)^n.$$

Substituting this in (1) above gives

$$r = \frac{1}{n} \log_2 (1+i)^n = \log_2 (1+i).$$

Thus r is a simple monotone function of the interest rate in the ordinary sense, and maximizing r is equivalent to maximizing i .

When n is large we expect the fraction of wins to be roughly p , i. e. $\frac{W}{n} \approx p$, while $\frac{L}{n} \approx q = 1-p$.

Thus

$$r_n \approx G = p \log(1+e) + q \log(1-e). \quad (4)$$

This statement can be made more precise by using the laws of large numbers.

According to the weak law of large numbers, given any two positive numbers

ϵ and δ a number N can be found such that if $n > N$, the probability is at least $1 - \epsilon$ that $|r - G| < \delta$. According to the strong law of large numbers, given any two positive numbers ϵ and δ a number N can be found such that the probability is at least $1 - \epsilon$ that $|r - G| < \delta$ after N bets and will remain so no matter how many more bets are made. An equivalent statement is that with probability one,

$$\lim_{n \rightarrow \infty} r_n = G$$

No matter which way you look at it, as the number of bets becomes very large, the gambler becomes more and more certain that his effective interest rate will be very close to G .

G is a function of e which has a maximum for some value of e . It is easily shown that the maximum occurs when $1+e=2p$, and hence $1-e=2q$. This gives

$$\begin{aligned} G_{\max} &= p \log 2p + q \log 2q = 1 + p \log p + q \log q \\ &= 1 - H(p) \end{aligned} \quad (5)$$

So that G_{\max} is equal to the rate of transmission over the channel by which the tips are received!

If one gambler bets always the optimum fraction of his holdings while a second bets a non-optimum fraction of his money on each bet, the effective interest rate for the first approaches G_{\max} with probability one while that for the second approaches some lower value. It follows that the probability approaches one as n approaches infinity that the first gambler will have more money than the second. The same result holds if the second gambler does not bet a constant fraction of his money on each bet as long as he deviates from the optimum by at least some fixed amount or at least a fixed fraction of the bets.

1. In information theory the problem often occurs of maximizing an expression of the form

$$\sum A_i \log x_i \quad (6)$$

by optimum choice of the x_i subject to the constraint that their sum is constant. The solution is that the x_i are proportional to the A_i .

In other words if one gambler bets according to the above scheme and a second according to any significantly different scheme, the probability approaches one as n approaches infinity that the first gambler will have more money than the second after n bets.

This is not to say that this method of betting is the only way a "rational" man would behave. While very persuasive in a general way, there are situations and systems of values or utilities which would lead to other methods of play, thus if the (remote) possibility of the extreme winning of $2^W V_0$ were sufficiently important (e. g. the only possible way to save the gambler's life) he would be well advised to bet maximum expectation (all on the most probable event).

Now consider the more general problem in which there are m events (outcomes of a horse race, for example) with probabilities P_1, P_2, \dots, P_m . The gambler receives a tip, one of n messages, which may not be reliable, perhaps because of noise in the communication channel. But the gambler is assumed to know how reliable the tips are by knowing the probability if event i occurred (or will occur) of tip j :

$p_i(j)$ = probability of tip j if event i occurs.

In addition to this the odds are assumed known

α_i = dollars returned per dollar bet if i occurs. The odds will be called fair if

$$P_i \alpha_i = 1, \quad (7)$$

and if the equality

$$\sum \frac{1}{\alpha_i} = 1 \quad (8)$$

holds, we shall say there is "no track take". (Note that "fair odds" implies "no track take" since, by (7) $1/\alpha_i = P_i$ and $\sum P_i = 1$.) "No track take" turns out to simplify the analysis greatly, since it permits covering bets with no loss, and hence makes betting all of one's holdings on every bet no less general than permitting holding back part of one's money. Note that if one bets $\frac{1}{\alpha_i}$ dollars on each event, he will have bet exactly one dollar and will have one dollar returned regardless of the outcome.

As an example, in pari mutual betting, the track takes a certain percent of all money bet and divides the rest among the people who bet on the winning horse. If the track takes t percent, and if n_i dollars are bet on the i th event, the odds are

$$\alpha_i = (1-t) \frac{\sum n_i}{n_i} \quad (9)$$

and

$$\sum \frac{1}{\alpha_i} = \frac{1}{1-t} \frac{\sum n_i}{\sum n_i} = \frac{1}{1-t} \quad (10)$$

If there is no track take, $t=0$, and

$$\sum \frac{1}{\alpha_i} = 1.$$

The gambler's strategy can be described by giving the percent of his holdings which he will bet on event i if he receives tip j . This will be denoted by $a(i/j)$.

First let us assume fair odds, which implies no track take. As was stated above, this means there is no loss of generality in assuming that the gambler bets all his holdings, since he can cover bets with no risk of loss. Then each bet multiplies his holdings by a factor $a(i/j)\alpha_i$ if event i occurs and he had received tip j . Suppose $W(i, j)$ denotes the number of times he received tip j and event i occurred in a total of n bets.

Then

$$V_n = \prod_{i,j} [a(i, j)\alpha_i]^{W(i, j)} V_0 \quad (11)$$

This gives an effective interest rate

$$r_n = \sum_{i,j} \frac{W(i, j)}{n} \log [a(i/j)\alpha_i] \quad (12)$$

which has as its limit with probability one,

$$G = \sum_{i,j} P_i p_j \log [a(i/j)\alpha_i] \quad (13)$$

The relationship between r and G is the same as in the simple case discussed first.

With fair odds, $\alpha_i = 1/P_i$, and hence,

$$G = \sum_{i,j} P_i p_j (j) \log a(i/j) - \sum_{i,j} P_i p_j (j) \log P_i \quad (14)$$

Summing on j first and noting that $\sum_j p_i(j)=1$, the last term becomes

$$-\sum_i P_i \log P_i = H(x) \quad (15)$$

Because of "no track take" we can assume that the gambler will bet all his money, i.e. we can assume the constraint $\sum_i a(i/j)=1$, and we can maximize separately the parts of the sum in (14) for each value of the index j . As before, (equation (6)), the $a(i/j)$ must be proportional to $P_i p_i(j)$. Since

$$\sum_i a(i/j)=1,$$

$$a(i/j) = \frac{P_i p_i(j)}{\sum_i P_i p_i(j)} \quad (16)$$

$$= \frac{p(i,j)}{Q(j)} = q_j(i)$$

where $p(i,j)$ is the probability that i occurred and tip j was received, $Q(j)$ is the probability of tip j , and $q_j(i)$ is the probability that event i occurred if tip j was received. Then

$$\begin{aligned} G_{\max} &= \sum p(i,j) \log q_j(i) + H(x) \\ &= H(x) - H_y(x) = R \end{aligned}$$

where x represents the event and y the tip. But again this is just the rate of transmission over the communication channel carrying the tip!

Now suppose that the odds are not necessarily fair, but that there is still no track take. The only change is that we cannot assume that $\sum_i p_i=1$, and hence the last term is $\sum p_i \log \alpha_i$ instead of $\sum p_i \log p_i$. Denoting this by $H(\alpha)$, G becomes

$$\begin{aligned} G &= -H_y(x) + H(\alpha) \\ &= -H_y(x) + H(x) + H(\alpha) - H(x) \\ &= R + R_0 \end{aligned}$$

where R is the rate of transmission of information and $R_0 = H(\alpha) - H(x)$. R_0 is independent of the tips, and hence we can see its significance by considering the case where the tips give no information. Then R_0 is the maximum effective interest rate possible with no tips. R_0 is greater than or equal to zero, and it equals zero only when $\frac{1}{\alpha_i} = p_i$, i.e. fair odds. R_0 represents the maximum effective interest rate achievable by taking advantage of the fact that the odds are not fair.

It is interesting to note that it is best to bet an amount of money $a(i/j)$ proportional to $q_j(i)$ regardless of the odds. One would think

that to take best advantage of unfair odds the bets should be adjusted differently for different odds, but this is not the case, at least for this type of betting.

Now assume a track take, i. e. $\sum \frac{1}{\alpha_i} < 1$. This case is considerably more difficult mathematically, so the results will only be outlined here. In general the gambler should hold back some money. Arrange the events in order of decreasing expectation (conditional on the available information), i. e. in order of goodness of the bets. At some point a line is drawn and bets placed only on the events above the line. Bets are made in proportion to the conditional probability of their occurrence, holding back some of the money. It turns out generally that some of the events bet on, the ones just above the line, have expectation less than one, i. e. $q_j(i) \alpha_i < 1$, even though such bets would seem to be quite poor.

How to Pay the Forecaster

The following analysis was considered by I. J. Good in England, and by Andy Gleason of Harvard University. The problem concerns piecework payment to a consultant for predictions, the payment to be made according to how good the prediction is.

Instead of the simple weather forecasts which are customarily made, use a more sophisticated system in which probabilities are given for each possible weather event. For example the weather man might say, "The probability is one-half that it will snow, one-sixth that it will rain, and one-third that it will be fair".

Now let us suppose that the client wishes to pay the forecaster day-by-day, and by merit. Thus it would seem that a relatively high fee should be paid if the forecaster assigns a high probability to the event which actually occurs, and a low fee should be paid if the forecaster assigns a low probability. But exactly what function of p ?

Now let us consider the forecaster's viewpoint. Let us suppose that he is more worried about how much money he will be paid than about good forecasting. Let us assume that the function of p , $f(p)$ which is his payment, is known to him (as part of his contract) and let us assume he knows the probabilities of the various events which he is attempting to forecast. Then he might attempt to optimize mathematically his payoff by reporting a number a_i as the probability of event i instead of its true probability p_i . His expected payoff in that case would be

$$P = \sum p_i f(a_i)$$

which he would maximize subject to the constraint $\sum a_i = 1$, since the a_i must look to the client like probabilities. Using the method of Lagrangian multipliers, we find that the a_i satisfy the equation

$$p_i f'(a_i) + \lambda = 0$$

for each value of i . These equations together with the constraint equation enable the forecaster to solve for the prediction a_i which will pay best.

Now, getting back to the client's viewpoint, he would like the prediction which he receives to equal the actual probability, i. e. $a_i = p_i$. This will be the case if and only if

$$p_i f^1(p_i) + \lambda = 0$$

for all p_i , or in other words if

$$x f^1(x) + \lambda = 0$$

The solution of this differential equation is

$$f(p) = -\lambda \log p + C$$

and if this is to be a maximum, the second derivative should be negative, or λ should be negative.

$$f(p) = A \log p + B \qquad A > 0$$

Now consider what the average payment is:

$$\begin{aligned} P_{ave} &= A \sum p_i \log p_i + B \\ &= B - A H(x) \end{aligned}$$

The forecaster is paid a fixed salary from which is deducted an amount proportional to the client's uncertainty about the predicted event after the prediction!

NOTES ON RELATION OF ERROR PROBABILITY TO DELAY IN A NOISY CHANNEL

Lecture by C. E. Shannon, August 30, 1956

The ordinary coding theorems assert something about what can be done in the limit of very long codes. They do not give information as to how long the code must be to approach within a certain tolerance of the limiting behavior. This question, the relation of probability of error and length of code, is of considerable interest. Results here bear about the same relation to earlier results as the central limit theorem in probability bears to the law of large numbers. In fact, at a key point in proving the theorems, the law of large numbers is used in the first case and a generalization of the central limit theorem in the second case.

The first type coding theorem relates to coding a source into binary digits (say). If the source produces letters at a regular rate and block coding is to be used a result may be obtained relating error probability (this is here the probability of rare sequences for which no binary sequences are available) and the rate at which binary digits are available. It is convenient to use a measure reliability, E , rather than probability of error directly.

$$E = \frac{1}{n} \log P_e^{-1}$$

where n is the block length and P_e the probability of error with best coding. As n increases, E approaches a limit in the case of sources described by a Markoff process. For the simplest case, that in which the language consists of a sequence of letters chosen independently from a finite alphabet with probability p_i for the i^{th} letter in the alphabet, the limiting E can be given in parametric form (parameter s) as follows.

Let $q_1(s) = p_1^{1-s} / \sum_1 p_1^{1-s}$. Then if $E(s)$ is the limiting reliability and $R(s)$ the rate of binary digits available for coding (per letter of text), we have

$$E(s) = \sum_1 q_1(s) \log \frac{p_1}{q_1(s)}$$

$$R(s) = \sum_1 q_1(s) \log q_1(s)^{-1}$$

A complete solution can also be given in the general Markoff case but is more involved.

The second type of coding theorem relates to coding a sequence (say) of binary digits into a noisy channel in such a way as to have a small probability of error after decoding. The problem involving delay in this case is to determine for a block length of code n and an input rate R the probability of error for the optimal code. We limit ourselves to discrete memoryless channels with finite alphabets. It is convenient also to use a reliability measure $E = \frac{1}{n} \log P_e^{-1}$.

The problem is that of estimating E as a function of R , or, as it turns out, E and R as functions of s . Upper and lower bounds are found on the probability of error for codes by a number of different arguments. The most powerful argument for showing the existence of codes is by the random coding procedure. Random codes are improved when the rate R is small by an expurgating procedure. This is the elimination of code words which are particularly close together. To establish lower bounds on the probability of error, the most powerful argument is by the sphere packing method. This is the generalized analog of arguments to the effect that one cannot get more than $\frac{V}{v}$ spheres of volume v in a room of volume V . The expurgated random code and the sphere packing argument

determine the asymptotic E exactly for rates R between a certain critical value and the channel capacity. In fact, as one approaches channel capacity the optimal probability of error for a given delay is more and more nearly determined. For rates below the critical rate, the bounds diverge. Another type of lower bound on probability of error, suggested by Elias for the binary symmetric channel, becomes more powerful in evaluating E . This is a bound based on the minimum separation between words in a code. It turns out that for rates near zero the probability of error is controlled chiefly by code words which are "close together".

U56

NOTES ON THE KELLY BETTING THEORY OF NOISY INFORMATION

Lecture by C. E. Shannon, August 31, 1956

[108]

In most communication studies the analysis stops when the message is received. No action based on the message is contemplated. John Kelly has considered a problem in which action is taken based on the received message, namely, the messages are assumed to be tips on the outcome of events and a gambler may place bets on these events. The problem is to determine the gambler's optimal system of betting and the value of the channel to him. It is assumed that the channel keeps operating and that the gambler can reinvest his winnings. If after n plays of this game the gambler has V_n dollars, we define his effective interest rate as $R = \frac{1}{n} \log_2 V_n / V_0$. We assume M events (e.g. entries in a horse race) with probabilities of occurrence P_1, P_2, \dots, P_M . The gambler receives a tip, one of n messages, which may not be reliable, but the gambler knows the probability $p_1(j)$ of tip j if event i occurs. The available odds for betting are α_1 dollars paid per dollar bet if i occurs. Odds are called fair if $P_1 \alpha_1 = 1$. We say there is no track take if $\sum_i \frac{1}{\alpha_i} = 1$. In the case of no track take, it is possible to effectively hold back a dollar by betting $\frac{1}{\alpha_i}$ dollars on event i for each i , since then one dollar is bet and one dollar always returned. Thus without loss of generality all the capital can be bet each time.

Assuming fair odds, (this implies no track take) it turns out that the expected interest rate is maximized if the gambler bets money on event i when tip j is received in proportion to $P_i p_1(j)$. When he bets this way his interest rate turns out to be

$$G = H(x) - H_y(x) = R$$

That is, his interest rate is the rate of transmission in communication theory over the channel carrying the tip. His interest rate is better than that of any gambler who deviates significantly from this strategy (with probability 1), that is, any gambler who does not bet this way a fraction of time $> \epsilon > 0$.

If there is no track take but the odds are not necessarily fair, it turns out that the best interest rate becomes

$$G = R + R_0$$

where R is the rate of transmission for the channel and R_0 is the maximum effective interest rate with no tips. It is the rate of interest one can obtain from the fact that the probabilities P_i are not equal to the betting odds $\frac{1}{a_i}$.

The situation is somewhat more complex when there is a track take.

Reference: John Kelly: "A New Interpretation of Information Rate",
Bell System Technical Journal, July, 1956.

The Fourth-Dimensional Twist

or

A Modest Proposal in Aid of the American Driver in England *

Claude E. Shannon

An American driving in England is confronted with a wild and dangerous world. The cars have the driver on the right and he is supposed to drive on the left side of the road. It is as though English driving is a left-handed version of the right-handed American system.

I can personally attest to the seriousness of this problem. Recently my wife and I, together with another couple on an extended visit to England, decided to jointly rent a car. Usually when we drove the men would sit in the front seat, the women in the back. With our long-ingrained driving habits the world seemed totally mad. Cars, bicycles and pedestrians would dart out from nowhere and we would always be looking in the wrong direction. The car was usually filled with curses from the men and with screams and hysterical laughter from the women as we careened from one narrow escape to another. The passengers were given to sudden involuntary motions - shielding the face or slamming on non-existent brakes. The turn indicator and windshield wiper controls were also reversed from American practice and we found ourselves signaling turns with the windshield wiper - fast for a right turn, slow for a left. The whole driving situation was not particularly improved by the narrowness of English streets and the high speed of English drivers. Nor was our inner security increased by the predilection of the English for building stone walls immediately adjacent to the roads.

This paper will develop a novel solution to this problem which

* This research was carried out in Trinity term, 1978 while the author was a Visiting Fellow at All Souls College, Oxford.

incidentally can also be used for the Englishman driving in America.

In Fig. 1 we see two triangles. They are congruent but one cannot be slid around in the plane to coincide with the other since one is, so to speak, a left-handed version of the other. A "flatlander", limited to living in the plane, could scarcely conceive how triangle A could be moved into coincidence with B, but we, as three-dimensional beings, easily understand rotating the triangle A about one of its sides and then sliding it into coincidence with the other.¹

In an analogous way, in three dimensions we often have right- and left-handed objects - a pair of gloves, for example, or an American car compared to an English car of the same type. If we had access to a fourth dimension, one could turn a left-handed glove 180° through the fourth dimension and it would reenter the third dimension as a right-handed glove.² This facility would be useful in many ways. Both shoemakers and screwmakers would benefit. The former would need only right-footed lasts, the latter only right-handed taps and dies. Left-handed children could be flipped through the fourth dimension to become right-handed, since the world of tools, writing, etc., is for the most part more friendly to the right-handed. Contrariwise, right-handed baseball pitchers might choose to become southpaws. Our American driver coming to England might choose to undergo this fourth-dimensional twist which would turn his perception of England from left-handed to right-handed.

Alas, no one has found a method to rotate an object through the fourth dimension. However, equally effective would be a rotation for our American driver of all of England through the fourth dimension. This concept no doubt sounds grandiose and utterly impractical - the

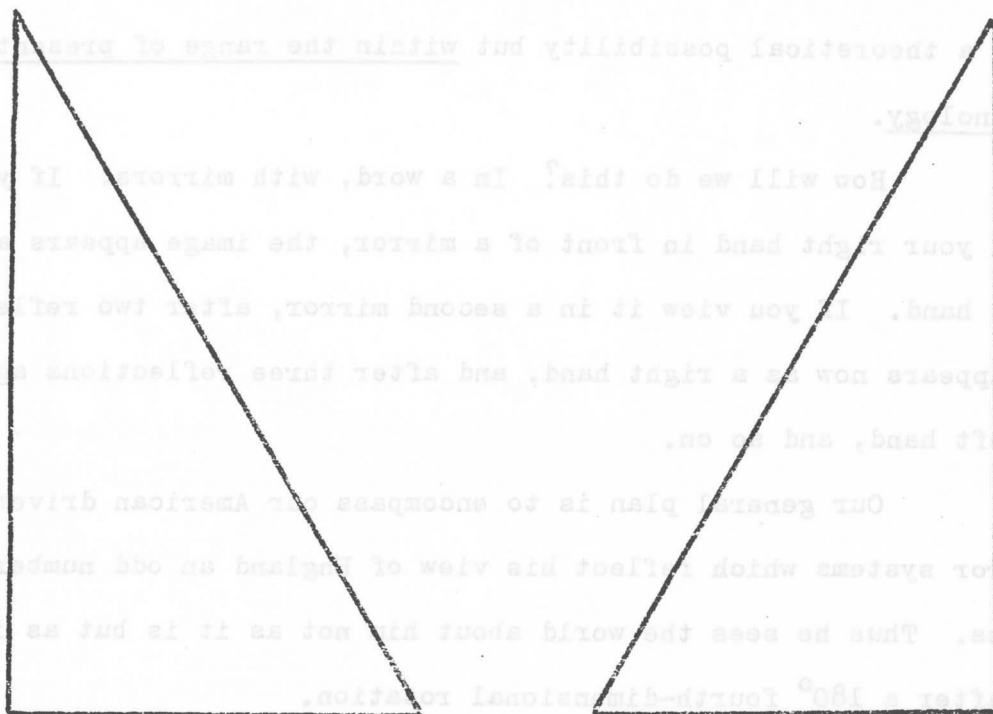


Fig. 1

idle dream of a mathematician - but we will show that it is not only a theoretical possibility but within the range of present-day technology.

How will we do this? In a word, with mirrors. If you hold your right hand in front of a mirror, the image appears as a left hand. If you view it in a second mirror, after two reflections it appears now as a right hand, and after three reflections again as a left hand, and so on.

Our general plan is to encompass our American driver with mirror systems which reflect his view of England an odd number of times. Thus he sees the world about him not as it is but as it would be after a 180° fourth-dimensional rotation.

To accomplish this we have two mirror systems. The side mirror system is shown in Fig. 2, where we see the driver, from the back, sitting in his English car. There are five mirrors in the car, two on his right, two on his left, and one above his head. These serve to reflect images from the left over his head and down again so they come in from the right.³ Similarly, light rays from the right are reflected over his head and down to come in from his left. Thus, if he turns his head to the right side of the page, he will see, by a triple reflection, an image of the object (an arrow) which is on the left of the page. In the same manner, if he looks to the left of the drawing, he will see what is on the right of the car.

To summarize, this group of five mirrors is so arranged that when he looks to his right he will see what is on his left - when he looks to his left he will see what is on his right.

Another set of mirrors provides for forward and backward vision. These are shown in Fig. 3, where we see the driver from above.

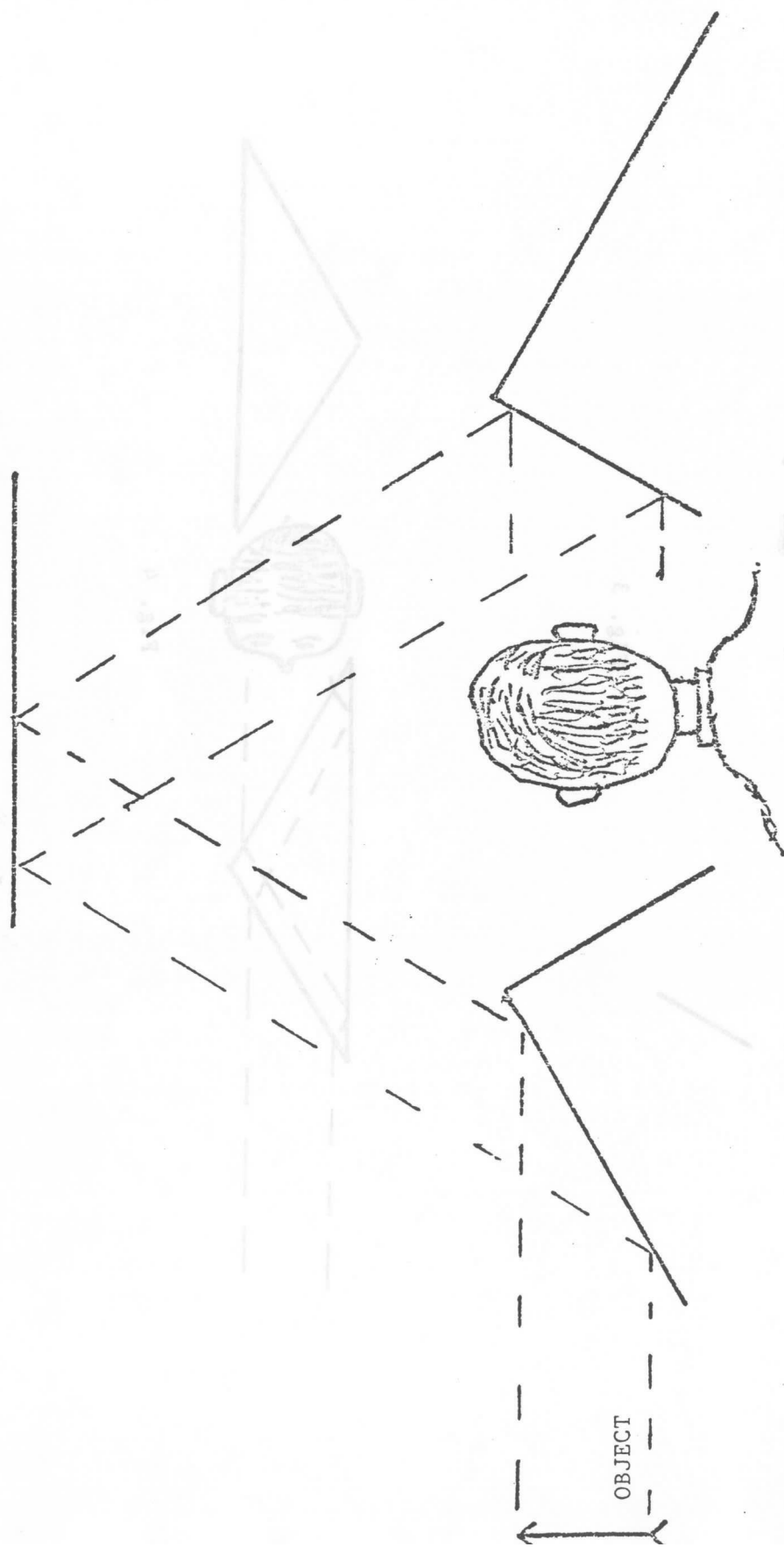


Fig. 2



REAR
VISION



Fig. 3



Fig. 4

For forward vision, three mirrors reflect the front visual field about a vertical axis.⁴ Some light rays are indicated by letters A and B to show how the interchange of right and left takes place. The object (the usual arrow) appears to the driver as a reversed image (again somewhat farther away because of the longer path).

A second set of three mirrors accommodates vision in the backward direction. If our driver should turn his head around, perhaps in driving in reverse or possibly to look at his passengers in the back seat, he will again see a left-right reversed image.

These four mirror systems totally encompass our American driver. Wherever he looks, he sees a reversed image of England - always reflected three times. For him, England has been rotated 180° through the fourth dimension!

A further detail must be accounted for here. The rear-vision mirror in an ordinary car corresponds to one reflection - in looking through it we see words reversed and, in fact, catch a tiny glimpse of the left-handed world we have been talking about. To keep our system consistent, and to keep our American driver comfortable, we have devised a rear-vision mirror using a double reflection, as shown in Fig. 3. The driver looks up and to the right, as he would in an American car, and sees out by a double reflection through the rear window. This gives him the only glimpse he has of the real "right-hand" world, since a double reflection preserves handedness.

In Fig. 5 we see from above a car fitted with the fourth-dimensional twister. The actual car as well as the actual English road and countryside, are shown in heavy solid lines. In reality, the car is parked on the left side of the road. Another car is forward to the right and the road turns sharply to the right. The driver's perception, however, because of his mirror system which

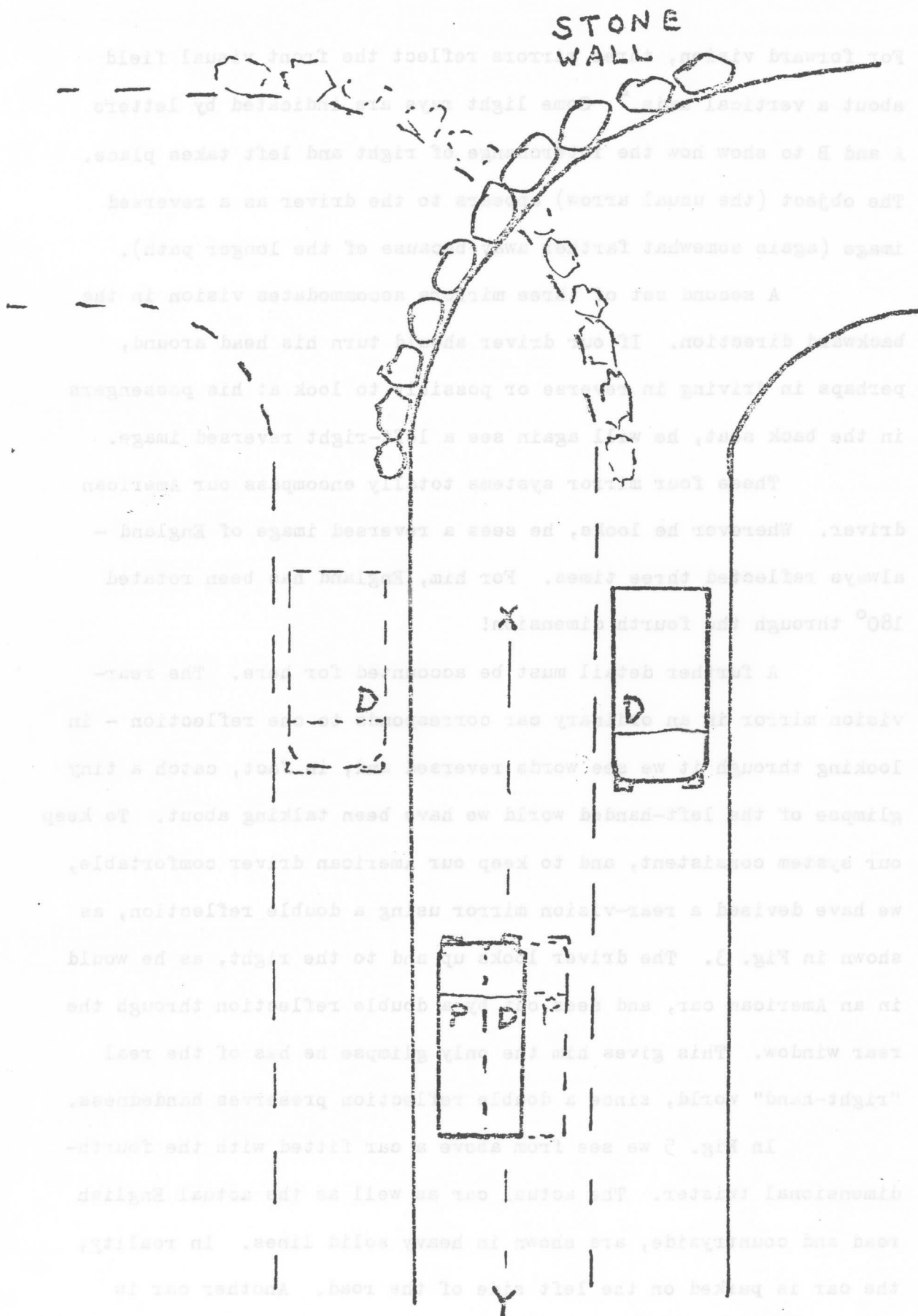


Fig. 5

reflects everything about the line XY, is that he is parked on the right side of the road, that the other car is at his left, and that the road turns sharply to the left. His perception of this situation is shown in dotted lines. Note that he even perceives his own car to have changed to an American car, and his passenger, P, on the front seat now appears to be on his right!

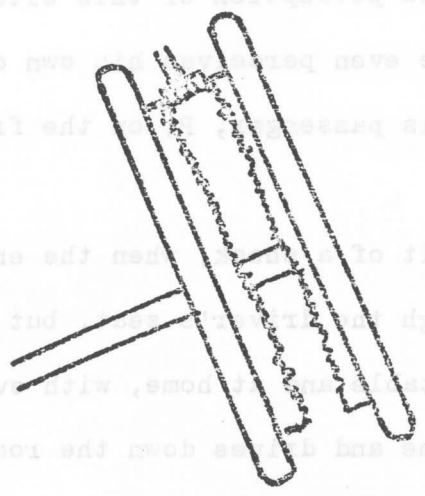
Entering this car may be a bit of a shock, when the entire world is reflected about a plane through the driver's seat, but after a moment our American will feel comfortable and at home, with everything as it "should be". He starts his engine and drives down the road. The road actually turns sharply to the right. In his perception of course it turns sharply to the left, so of course he turns to the left, directly into the stone wall, and is instantly killed.

This, of course, is what would have happened had we not foreseen his natural reactions to a reversed perception of the world. One must reverse not only the sensory input but also the motor output. Fig. 6 shows an attachment to the steering wheel which reverses its operation. When turned to the right, the vehicle actually turns to the left and vice versa. This operates much as differential gears in automobiles.

With this addition our American driver will perceive a curve to the left and, in natural response, turn to the left. In fact the curve will be to the right and the mechanism will reverse his intent and turn the car to the right.

This, then, is the basic idea of the fourth-dimensional twist. There are, however, some loose ends to be dealt with. The perceptive reader may wonder about road signs. Our American driver, viewing everything through a triple reflection, sees all of the road signs

reflects everything about the line XY, is that he is parked on the right side of the road, that the other car is at his left, and that the road turns sharply to the left. His perception of this situation is shown in dotted lines. Note that he even perceives his own car to have changed to an American car, and his passenger seat now appears to be on his right!



Entering this car may be a bit of a shock, but the entire world is reflected about a plane through the car, but after a moment our American will feel comfortable at home, with everything as it "should be". He starts his engine and drives down the road. The road actually turns sharply to the right. In his perception of course it turns sharply to the left, so of course he turns to the left, directly into the stone wall, and is instantly killed.

Fig. 6

This, of course, is what would have happened had we not foreseen his natural reaction to a reversed perception of the world. One must reverse not only the sensory input but also the motor output. Fig. 6 shows an attachment to the steering wheel which reverses its operation. When turned to the right, the vehicle actually turns to the left and vice versa. This operation is such as differential gears in automobiles. With this addition our American driver will perceive a curve to the left and, in natural response, turn to the left. In fact the curve will be to the right and the mechanism will reverse his intent and turn the car to the right.

Thus, then, is the basic idea of the fourth-dimensional twist. There are, however, some loose ends to be dealt with. The perceptive reader may wonder about road signs. Our American driver, viewing everything through a triple reflection, sees all of the road signs

in reverse, as, for example, in Fig. 7. How is he to find his way about? The answer is ridiculously simple. We have already pointed out that his rear-vision mirror gives a double reflection and hence a normal view of the real world. All he need do is back his car up to the road sign and read it through his rear-vision mirror!

A more troublesome problem is that of centrifugal force. In the situation of Fig. 5, our driver is actually turning to the right but perceives himself to be turning to the left. Centrifugal force will opt for actuality. Our driver will, surprisingly, find himself driven to the inside of the curve rather than the outside, a most uncomfortable and confusing sensation.

To solve this problem, the reversal of centrifugal force, might seem as impossible as the twist of England through the fourth dimension. After all, centrifugal force is given by the formula

$$f = m \frac{\omega^2}{R}$$

A radius R of course is always positive, ω^2 as a square is necessarily positive, and surely a mass m must be positive, so how can we arrange for the centrifugal force f to be negative? Like Columbus and the egg, the answer is very simple when given. If we immerse the mass in a liquid of higher density, it acts as though it, itself, had a negative mass. The liquid itself presses the object in the direction of acceleration!

This concept is shown in Fig. 8. Our driver is now enclosed in a scuba-diving suit within a compartment which is filled with a liquid having a specific gravity of approximately 2. Of course he would tend to rise in this liquid but he is held down firmly by his seatbelt. A snorkel provides for his breathing and altogether, with our various devices, he feels very much as though he were at home in America!

in reverse, as, for example, in Fig. 7. Now is he to find his way about? The answer is ridiculously simple. We have already pointed out that his rear-vision mirror gives a double reflection and hence a normal view of the real world. All he need do is look his car up to the rear sign and read it through his rear-vision mirror!

A more troublesome problem is that of centrifugal force. In the situation of a car turning to the right but perceiving the right but perceiving the left. Centrifugal force will act for actuality. The driver will, surprisingly, find himself driven to the inside of the curve rather than the outside.

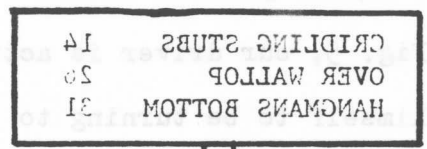


Fig. 7

To solve this problem, the reversal of centrifugal force, might seem as impossible as the twist of England through the fourth dimension. After all, centrifugal force is given by the formula

$$F = \frac{mv^2}{R}$$

A radius R of course is always positive, as is a square is necessarily positive, and surely a mass m must be positive, so how can we arrange for the centrifugal force F to be negative? Like Columbus and the egg, the answer is very simple when given. If we increase the mass in a liquid of higher density, it acts as though it, itself, had a negative mass. The liquid itself presses the object in the direction of

acceleration!

This concept is shown in Fig. 8. Our driver is now enclosed in a scuba-diving suit within a compartment which is filled with a liquid having a specific gravity of approximately 2. Of course he would tend to rise in this liquid but he is held down firmly by his seatbelt. A snorkel provides for his breathing and altogether, with our various devices, he feels very much as though he were at home in

water!

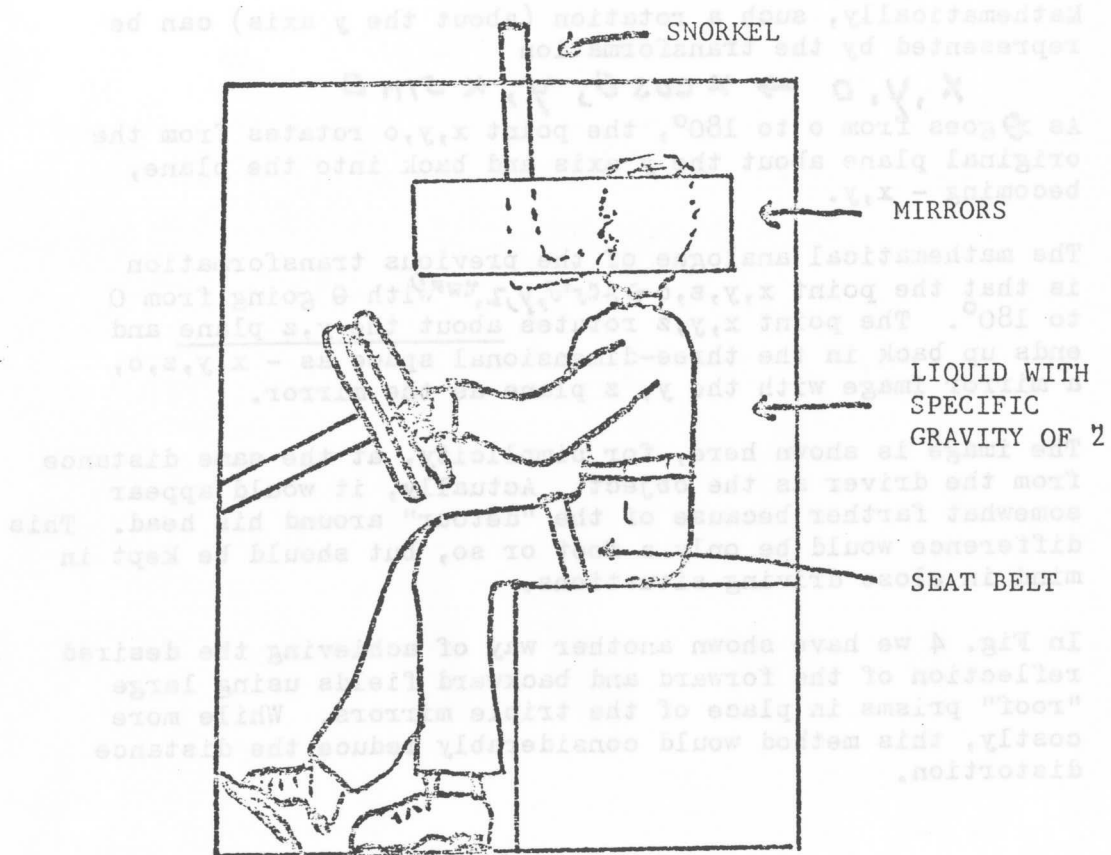


Fig. 8

FOOTNOTES:

1. Mathematically, such a rotation (about the y axis) can be represented by the transformation

$$x, y, 0 \rightarrow x \cos \theta, y, x \sin \theta$$

As θ goes from 0 to 180° , the point $x, y, 0$ rotates from the original plane about the y axis and back into the plane, becoming $-x, y$.

2. The mathematical analogue of the previous transformation is that the point $x, y, z, 0 \rightarrow x \cos \theta, y, z \sin \theta$ with θ going from 0 to 180° . The point x, y, z rotates about the y, z plane and ends up back in the three-dimensional space as $-x, y, z, 0$, a mirror image with the y, z plane as the mirror.
3. The image is shown here, for simplicity, at the same distance from the driver as the object. Actually, it would appear somewhat farther because of the "detour" around his head. This difference would be only a foot or so, but should be kept in mind in close driving situations.
4. In Fig. 4 we have shown another way of achieving the desired reflection of the forward and backward fields using large "roof" prisms in place of the triple mirrors. While more costly, this method would considerably reduce the distance distortion.

A Rubric on Rubik Cubics⁽¹⁾

Claude E. Shannon

Once puzzledom was laissez faire
With rebus, crosswords, solitaire.
Comes now the Rubik Magic Cube
For Ph. D. or country rube.
This fiendish clever engineer
Entrapped the music of the sphere.
It's sphere on sphere in all 3 D –
A kinematic symphony!

Ta! Ra! Ra! Boom De Ay!
One thousand bucks a day.
That's Rubik's cubic pay.
He drives a Chevrolet.⁽²⁾

Forty-three quintillion plus⁽³⁾
Problems Rubik posed for us.
Numbers of this awesome kind
Boggle even Sagan's mind.⁽⁴⁾
Some chaps pry their cubes apart
Then reassemble to the "start".

Not cricket! A rude game's afoot

And up with which we will not put!

Ta! Ra! Ra! Boom De Ay!

Cu-bies in disarray?

First twist them that-a-way,

Then turn them this-a-way.

Respect your cube and keep it clean.

Lube your cube with Vaseline.

Beware the dreaded cuber's thumb,

The callused hand and fingers numb.⁽⁵⁾

No borrower nor lender be.

Rude folk might switch two tabs on thee,

The most unkindest switch of all,

Into insolubility.⁽⁶⁾

In-sol-u-bility.

The cruelest place to be.⁽⁷⁾

However you persist

Solutions don't exist.

While most folk watch the idiot tube

Cubemeisters spin the Rubik cube.

Minh Tai's the champ — he's fast as sin.

Minh solves *his* cube in half a min.⁽⁸⁾

John Conway leads a Cambridge pack

And solves *his* cube behind his back.⁽⁹⁾

Singmaster write THE BOOK — first rank;

Now cubes while riding to the Bank.⁽¹⁰⁾

Here now a heavyweight!

Programming potentate!

Software sophisticate!

Morwen B. Thistlethwaite!⁽¹¹⁾

Eschewing this dull 3 D place

Joe Buhler cubes in hyperspace.⁽¹²⁾

All hail Dame Kathleen Ollerenshaw,

A mayor with fast cubic draw.⁽¹³⁾

Is cubing just a crashing bore?

Let Talken's robot do this chore.⁽¹⁴⁾

God moves in geodesic ways

And solves His cube in twenty plays.⁽¹⁵⁾

Cubemeisters one and all,

Their cubes find final rest

Bronzed in the Hall of Fame

In lovely Budapest.

The battle's joined in steely grip:

Man's mind against computer chip,
With theorems wrought by Conway's eight
'Gainst programs writ by Thistlethwaite.
Can multi-billion neuron brains
Beat multi-megabyte machines?
The thrust of this theistic schism —
To ferret out God's algorism!

CODA:

He (hooked on cubing) with great enthusiasm:

Ta! Ra! Ra! Boom De Ay!
Men's schemes gang aft agley.
Let's cube our life away!

She: Long pause (having been here before):

-----OY VAY!

- (1) When T. S. Eliot published "The Waste Land" in 1922 with a wealth of footnotes, there was considerable commotion among the critics — should a work of art stand on its own feet or refer to such weighty tomes as *The Golden Bough*. The ambiguity, obscurity and even prurience of modern poetry are also under attack. We intend this to be clean as a hound's tooth, crystal clear, sensible as a dictionary, and with footnotes galore.

First off, this may be either read as a poem or, better, sung to "Ta! Ra! Ra! Boom De Ay!" (with an eight bar chorus). The verses should be sung solo, in a slightly bitter sardonic manner, a la Noel Coward or Bea Lillie; the choruses, in contrast, a joyous rousing salute to the cube.

- (2) A little poetic license here — the Wall Street Journal, Sept. 23, 1981, reports Rubik as receiving \$30,000 a month from cubic royalties, but driving a "run-down rattling Polski Fiat". This would neither scan nor rhyme as well as Chevrolet.

- (3) There are

$$\frac{8!}{2} \cdot \frac{12!}{3} \cdot \frac{3^8}{3} \cdot \frac{2^{12}}{2} = 43252\ 00327\ 44898\ 56000$$

possible arrangements of the cube.

- (4) It would take *billions and billions* of "billions and billions" for forty-three quintillion plus.
- (5) While not as debilitating as weaver's bottom or hooker's elbow, cuber's thumb can be both painful and frustrating. For more on these occupational ailments see recent issues of "The New England Journal of Medicine".
- (6) A friend of mine, Pete, an expert cuber, told me of encountering a friend Bill at a hobby shop. Bill gave Pete his cube, saying that he had been working for days without success. After a few minutes, Pete turned it into a position where he could see that two tabs had been interchanged.

Pete: Bill, somebody has switched two tabs on your cube.

Bill: That's impossible. I've always carried it, or left it in my apartment, and nobody has keys to get in there.

Pete: Nobody?

Bill: That's right, nobody. Just me and my girlfriend.

- (7) Especially in April.
- (8) Minh Tai, World Speed Champion, in a public demonstration solved six scrambled cubes, each in less than 30 seconds.
- (9) Actually, he peeks a little. John Conway, the great Cambridge combinatorialist, in addition to his tour de force blindfold cubing has, with his colleagues, contributed much to Rubik cube theory.
- (10) Singmaster, David. *Notes on Rubik's Magic Cube*, now in its sixth edition.
- (11) A pioneer in programming computers to solve the cube. His program solves the cube in 52 or fewer moves.
- (12) Group theorist Buhler and his colleagues have developed a theory of higher dimensional cubes.
- (13) Renaissance woman, sometime mayor of Manchester, recreational mathematician, expert cubist and discoverer of the cubist thumb syndrome and its relation to the fetlock problem in horses.
- (14) In October 1981 the writer foresaw the need for a cubing machine and sketched the design of a pair of mechanical hands to be connected to a computer and manipulate a cube. In the summer of 1982 a crack team of one M.I.T. student was assembled. Late in July the hands were making their first fumbling attempts to hold and manipulate a cube, when we received a crushing newspaper clipping from a friend. It seems that Dan Talken had assembled a crack team of Southern Illinois University students and beat us to the punch. My friend wrote one word across the slipping: "Scooped!"
- (15) Or so Singmaster finds it tempting to conjecture.

